

# The European Lake Microbiome: A Study in Complexity

**DISSERTATION**  
(KUMULATIV)

ZUR ERLANGUNG DES GRADES EINES  
DOKTOR DER NATURWISSENSCHAFTEN  
(DR. RER. NAT.)

VORGELEGT VON  
**THEODOR SPERLEA**  
AUS BIETIGHEIM-BISSINGEN

MARBURG, 2021



Die vorliegende Dissertation wurde von März 2017 bis Juni 2021 am Fachbereich Mathematik und Informatik unter Leitung von Prof. Dr. Dominik Heider angefertigt.

Vom Fachbereich Biologie der Philipps-Universität Marburg (Hochschulkennziffer 1180) als Dissertation angenommen am

Erstgutachterin: Prof. Dr. Anke Becker  
Zweitgutachter: Prof. Dr. Dominik Heider  
Weitere Mitglieder der Prüfungskommission:  
Prof. Dr. Hans-Ulrich Mösch  
Prof. Dr. Bernhard Seeger

Tag der Disputation: \_\_\_\_\_





# Erklärung

Ich versichere, dass ich meine Dissertation mit dem Titel “The European Lake Microbiome: A Study in Complexity” selbstständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen und Hilfsmittel bedient habe.

Diese Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den 7.7.2021

Theodor Sperlea



# The European Lake Microbiome: A Study in Complexity

## ABSTRACT

While it is known that microbes play many indispensable roles in ecosystems, the relationship between microbiomes and their environment is far from being well-understood. In part, this is the case because the methods necessary for studying environmental microbiomes, such as Next-Generation Sequencing and high-dimensional Machine Learning, have been developed relatively recently. However, the complex nature of ecosystems and environmental microbiomes acts as a further barrier to progress in this field of research.

This thesis develops methods and concepts used to gain insight into the ecology of microbiomes in lakes. It is based around two metabarcoding datasets sampled from lakes in Austria and the whole of Europe, respectively, and attempts to elucidate the microbiome's relationship to environmental parameters. To this end, a tool for GPS-based dataset enhancement and a machine learning framework for measuring microbiome covariation is developed. Building on this, the latent structure of the microbiome is estimated. In the discussion, a novel theory of information transmission in complex environments is described.

Taken together, the work included herein presents a thorough analysis of the European lake microbiome that takes the complexity of the study object into account. The results point towards parameters that act as drivers of lake microbiome structure as well as microorganisms that might act as keystone species for ecosystem functioning. Furthermore, this work might provide the basis for considerable future progress in the study of environmental microbiomes.

# The European Lake Microbiome: A Study in Complexity

## ZUSAMMENFASSUNG

Obwohl bekannt ist, dass Mikroben viele essenzielle Rollen in Ökosystemen spielen, ist die Beziehung zwischen Mikrobiomen und ihrer Umwelt noch recht unerforscht. Das liegt zum Teil daran, dass die für die Untersuchung von Umweltmikrobiomen notwendigen Methoden, wie Next-Generation Sequencing und hochdimensionales maschinelles Lernen, erst vor relativ kurzer Zeit entwickelt wurden. Die Komplexität von Ökosystemen wie auch Umweltmikrobiomen stellt jedoch ein weiteres Hindernis für den Fortschritt in diesem Forschungsgebiet dar.

Diese Arbeit entwickelt Methoden und Konzepte, um Einblicke in die Ökologie von Mikrobiomen in Seen zu gewinnen. Sie basiert auf zwei Metabarcoding-Datensätzen, die aus Seen in Österreich bzw. ganz Europa entnommen wurden, und versucht, die Beziehung des Mikrobioms zu Umweltparametern aufzuklären. Zu diesem Zweck wird ein Werkzeug zur GPS-basierten Datensatzanreicherung und ein maschinelles Lernverfahren zur Messung der Kovariation zwischen Mikrobiom und Umweltparametern entwickelt. Auf die damit generierten Resultate aufbauend wird die latente Struktur des Mikrobioms geschätzt. In der Diskussion wird eine neuartige Theorie der Informationsübertragung in komplexen Umgebungen beschrieben.

Insgesamt stellt die vorliegende Arbeit eine gründliche Analyse des europäischen Seenmikrobioms dar, die der Komplexität des Untersuchungsobjekts Rechnung trägt. Die Ergebnisse weisen auf Parameter hin, die als Treiber für die Struktur des Seenmikrobioms fungieren, sowie auf Mikroorganismen, die als Schlüsselspezies für das Funktionieren des Ökosystems fungieren könnten. Darüber hinaus könnte diese Arbeit die Grundlage für erhebliche zukünftige Fortschritte bei der Untersuchung von Umweltmikrobiomen bilden.

# Acknowledgments

I want to thank all people that, knowingly or unknowingly, have supported me in writing this thesis. First and foremost, I want to thank my wife Johanna Sperlea, who always had had my back as well as my parents, without whom I would not be at this point. Major thanks go to Dominik Heider, who has supervised me throughout this project, Anke Becker, who gladly stepped in as Erstprüferin, and Georges Hattab, who served as a sparring partner whenever I needed one. Furthermore, I want to thank Torsten Waldminghaus for feeding my scientific hunger and Franziska Löchel, Roman Martin and Marius Welzel, who became companions along the way. Furthermore, I need to thank Jens Boenigk and Daniela Beisser whose work and advice has been indispensable for the work presented here. I feel thankful for all the students and collaborators who contributed to my research, the examiners of this thesis for their time, and my friends for the much-needed distraction. Further thanks go out to Nils Richber and Isidore Nabi, who, in their ways, have showed me that straying from the path of one's scientific discipline does not necessarily make one go astray. I am deeply in debt to Alexandra Elbakyan and her work for open science, without this would simply not have been possible. Finally, I want to thank the Agentur für Arbeit Marburg and the current pandemic for granting me the calm and absence of distractions that was instrumental for me finishing this thesis.



# Contents

ABSTRACT	vii
ACKNOWLEDGMENTS	ix
<b>I INTRODUCTION</b>	<b>3</b>
1.1 The Experiment in Microbiology . . . . .	3
1.2 Complex Adaptive Systems in Ecology . . . . .	6
1.2.1 Complex Adaptive Systems . . . . .	6
1.2.2 Lake Ecosystems as Systems . . . . .	9
1.2.3 Environmental Microbiomes . . . . .	12
1.2.4 Sequencing and Barcoding of eDNA . . . . .	14
1.3 Computational Models for Microbiomes . . . . .	19
1.3.1 General Remarks . . . . .	19
1.3.2 Networks Models . . . . .	19
1.3.3 Ecological Stability and Biomonitoring . . . . .	22
1.3.4 Machine Learning Models . . . . .	24
1.3.4.1 Linear Regression Models . . . . .	26
1.3.4.2 Support Vector Regression . . . . .	28
1.3.4.3 Decision Trees and Ensembles of Decision Trees . . . . .	30
1.3.4.4 Feature Selection . . . . .	33
1.4 Details of This Work . . . . .	36
1.4.1 Datasets . . . . .	36
1.4.2 Assumptions of this work . . . . .	36
<b>2 PUBLICATIONS</b>	<b>41</b>
2.1 Publication I . . . . .	41
2.2 Publication II . . . . .	55
2.3 Publication III . . . . .	71
2.4 Publications not included in this thesis . . . . .	98
<b>3 DISCUSSION</b>	<b>99</b>
3.1 Introductory remarks . . . . .	99

3.2	The Comparability of Microbiomes . . . . .	100
3.3	A Theory for Microbial Biomonitoring . . . . .	102
3.3.1	The Microbiome as a Biosensor . . . . .	103
3.3.2	Information theory . . . . .	104
3.3.3	An Extended View of Information . . . . .	107
3.3.4	The Sensor and its Umwelt . . . . .	111
3.4	The Covariation Framework: Coupling and Complexity . . . . .	115
3.5	The Future of the Microbiome . . . . .	118
REFERENCES		153
APPENDIX A CURRICULUM VITAE		155



# 1

## Introduction

### 1.1 THE EXPERIMENT IN MICROBIOLOGY

Arguably, the scientific enterprise of microbiology came into being when Antonie van Leeuwenhoek, in the year 1677, wrote a letter to the Royal Society in London describing the *animalcules* he observed in drops of water using a microscope<sup>232,302</sup>. Two centuries later, Louis Pasteur, Robert Koch, and the other pioneers of microbiology discovered how to create and maintain pure cultures of specific microbial species<sup>45</sup>. What followed was an explosion of scientific inquiry into the microbial world, fuelled, in part by the institutionalization of microbiological laboratories and, in part, by the experimental method itself.

At its core, an experiment consists of a directed intervention into or manipulation of a study object and recording its response, conducted in such a way as to be sufficiently reproducible<sup>123,249</sup>. The optimal experimental setup controls all variables and parameters relevant to the study object except for those that are manipulated or recorded<sup>240</sup>. This is usually achieved by excising the study object from its original environment or breaking larger objects down into their constituent parts<sup>244,251</sup>. The degree of control of the study object exercised by the experimental setup allows researchers to test and falsify hypotheses regarding the study object, leading to the identification

and mechanistic explanation of correlative as well as causal relationships between the intervention and the study object's response<sup>123,322</sup>. The experimental method has paved the way for almost all progress in the natural sciences making it almost synonymous with the practise of science as such. Metaphorically speaking, the experimental method has turned the research scientist from a passive disciple of nature into an incessant interrogator, who can state hypotheses as inquiries that draw out the study object's testimony<sup>158,243</sup>.

In preparing a study object for an experiment, the scientist implicitly makes the assumptions that the properties of the object of interest are (i) not significantly determined by its original environment and (ii) explainable based on the properties of its parts when observed in separation<sup>109</sup>. While these assumptions usually hold for study objects that are mechanic in nature<sup>86</sup>, the same is not true for most objects that are of interest to the microbiologist as these are complex systems<sup>38</sup>. Biological objects are tightly interconnected to their environment and derive part of their functionality from this<sup>16,200,236</sup>. Similarly, the isolation of parts of a biological object from the whole or a biological object from its environment can lead to stark alterations in the object's or part's behavior<sup>295,306</sup>. This leads to an impasse the experimental microbiologist is well aware of: For instance, the *Escherichia coli* strains that have resided in laboratories for decades might behave completely different than the ones living in their natural habitat, the gut of a host organism. Nevertheless, studying this model organism in an experimental setting is the most direct path to new insights into the life of this bacterium.

It seems, however, that the development of the disparate set of methods somewhat laconically called 'omics methods has changed the situation and is now boosting a new generation of observational studies of complex systems<sup>1,270,318</sup>. Each of the 'omics methods can record a snapshot of the respective -ome, i.e., all the instances of a certain class in a given study object (for example, the genome contains all the genes, the proteome all the proteins, the lipidome all the lipids and the metabolome all the metabolites in an organism, a cell or tissue type)<sup>10,28,106,214</sup>. As such, 'omics methods drastically reduce the necessity to physically extract parts from wholes by enabling the observation of the former in a highly parallel manner. Furthermore, some 'omics methods can be extended to samples of multiple cells, organisms, *etc.* (and are, then, somewhat consequently called *meta-omics* methods), and applied to environmental samples, enabling researchers to study complex systems in their natural environment. At the same time, the vastness of the datasets generated when using 'omics methods affects the scientific practice. For one, it introduces the

necessity for time-efficient computational analyses. In fact, the analysis of the large datasets generated using 'omics methods is one of the main drivers for progress in bioinformatics, albeit not its sole *raison d'être*<sup>105</sup>. This goes hand-in-hand with a shift away from a knowledge-driven to a data-driven mode of doing science, as expertise derived from separate cases cannot easily be transferred to the high number of interdependent objects present in -omes<sup>161</sup>. Thus, it is no surprise that the arrival of 'omics methods rendered new study objects accessible, and with it, lead to the emergence of novel sub-fields, most of which are, currently, in a descriptive state<sup>28,76,227,283</sup>. One of these is the study of environmental microbiomes as an interdisciplinary study object of microbiology and ecology.

In this thesis, I present the methods used and the results generated in studying the relationship between the lake microbiome and its environment. To this end, I analyzed a dataset generated in a large-scale metabarcoding sampling effort that included lakes across Europe. Because this dataset does not stem from an experimental setting, an underlying question of this thesis is how to analyze environmental microbiomes while taking their complexity into account. To lay the groundwork for this work, I will first sketch out a basic theory of complex adaptive systems in section 1.2.1. By applying this theory to lake ecosystems, I will then define many of the terms most important for this thesis (section 1.2.2). After that, I will operationalize the most central term of this thesis, the (environmental) microbiome, in section 1.2.3. Because this operationalization is intricately linked to the identification of environmental microbiomes using DNA sequencing, this will be described in section 1.2.4. Section 1.3 examines methodical approaches developed by theoretical and computational ecology with regard to their use for the analysis at hand. A special focus is put on a set of machine learning methods and whether they might be useful when studying environmental microbiomes as complex systems (section 1.3.4). Finally, in section 1.4, I describe the datasets used in this thesis as well as the assumptions necessary to be able to work with the data at hand.

## 1.2 COMPLEX ADAPTIVE SYSTEMS IN ECOLOGY

### 1.2.1 COMPLEX ADAPTIVE SYSTEMS

Throughout this thesis, I will use some terms in a fairly specific sense that, colloquially, have a much broader meaning. One of these is the notion of the system, with which, unless explicitly otherwise noted, I will refer to a complex adaptive system. An object is a complex adaptive system if its properties emerge through the interactions among its parts as well as the object's interaction with its environment<sup>171,191,266</sup>. Because of this emergence, the system has properties not present in its parts, or, more colloquially, it is “more than the sum of its parts”<sup>16,53,61,118</sup>. While the parts of a system are intricately interconnected, they do not merge into the whole but remain distinct<sup>109,180,183,205</sup>. Along similar lines, complex adaptive systems can be characterized as consisting of a number of parts high enough so that not all possible interactions and relations between them can be realized<sup>191</sup>.

Systems can be distinguished from assemblages by the way they are constituted: In contrast to a system, which creates the distinction between itself and its environment through its organization, the assemblage is defined by an observer, usually as a statistical unit of entities with sufficiently high similarity<sup>118,151</sup>. Nevertheless, because systems are complex themselves and assemblages (usually) consist of complex systems, both inhabit a world of complexity – in contrast to experimental study objects insulated by the experimental setup. Note that, following this definition, complexity is not a gradual but a binary property: An object is either complex (and then I refer to it as system) or it is not. In this context, controlling or reducing complexity can only mean that a complex system is turned into a trivial, i.e., experimental, object.

A system decomposes if it is unable to reproduce its organization continuously, i.e., the interactions between its parts and, if necessary or possible, the parts themselves. The reader might notice that this statement is a tautology, yet it has a few interesting consequences<sup>136</sup>. First, the circular, self-reproducing self-relation, often termed autopoiesis, defines the boundary separating the system itself and its environment<sup>151,199,303</sup>: Everything that is reproduced by the system is part of the system, which makes systems operationally closed. Second, to maintain its organization, a system needs to maintain a lower level of entropy than is present in its environment<sup>242,274,310</sup>. Theoretical considerations in thermodynamics suggest that systems can avoid the thermodynamic equilibrium by (i) feeding on entropy and energy gradients in their environment<sup>155,308</sup> and (ii)

dissipating internally generated entropy<sup>108,215,243</sup>. Third, to do so, systems need to adapt their organizational structure because this defines the set of interactions a system can partake in<sup>178,191,199,237</sup>. Through adaptation and over time, the organization of the system and the structure of its environment is becoming increasingly interlinked. This, in turn, leads to the paradoxical situation that systems are distinct from their environment yet wholly dependent on and only understandable in the context of their respective environment<sup>310,326</sup>. Fourth and last, the interlinkage between the system and the environment creates sets of states, so-called attractors, in which the dynamics of the system are stable. Attractors are a strong indication that systems are chaotic, which would entail a high sensitivity to their environment's state, and thus the possibility that small perturbations lead to non-linear responses or bifurcations in the system's development<sup>113,128,178,242</sup>. As such, systems are at the “edge of chaos”, surrounded by either too little entropy to survive or by so much as to be in danger of disruption<sup>178,200</sup>.

Although the above abstract description applies to social, cognitive, and, as we will see, ecological systems, let me exemplify the properties of complex adaptive systems with a single prokaryotic cell. The cell consists of a high number of parts, i.e., lipids, proteins, carbohydrates, and nucleic acids, among other things, but it is only alive insofar as these parts interact<sup>260</sup>. In order not to decompose, the cell's metabolism has to keep going, which requires a steady inflow of energy and outflow of low-energy waste products. The processes essential to keep the cell alive continuously define the organizational boundary of the microbe, which makes the microbe an autopoietic system<sup>81,199</sup>. To extract enough nutrients and energy from its environment, an organism needs to be adapted to it – which is facilitated both by Darwinian evolution and adaptation through gene regulation and cellular feedback loops<sup>12,19,164</sup>. The structural and functional coupling arising from this is most obvious *ex negativo*, when, e.g., organisms cannot be cultured in a certain medium or require a long acclimation phase to re-adapt. Finally, because a microbial cell is organizationally closed, it can only register signals that induce a process in the cell. Signals that do not, e.g., bind to a receptor or cause physical changes to the cell wall will simply not have a relevant effect on the microbe.

The fact that the parts of systems can themselves be systems immediately suggests hierarchies of complex systems<sup>85,282</sup>. In most descriptions of such hierarchies, each level of organization contains comparable types of systems and interactions that differ from those at other levels<sup>36,61</sup>. One of the most popular hierarchies is the one presented in table 1.1 that connects the level of atoms

---

system or assemblage
Earth system
Biosphere
Biome
Landscape
Ecosystem
Community
Population
Organism
Cells/Organelles
Molecule
Atom

---

**Table 1.1:** A monolithic hierarchy of systems (and assemblages) in the living world. One would need to show that the hierarchy extends upwards (towards the Universe as a whole) and downwards (towards sub-atomic particles, quantum fields and, possibly, strings) if one were to argue that a monolithic hierarchy of systems is a formative concept in nature. However, as noted in the main text, there are many reasons to doubt this, and this is not the right place to make or debate such a claim. Nevertheless, this partial hierarchy should suffice as an representation of the underlying concept of levels of organization.

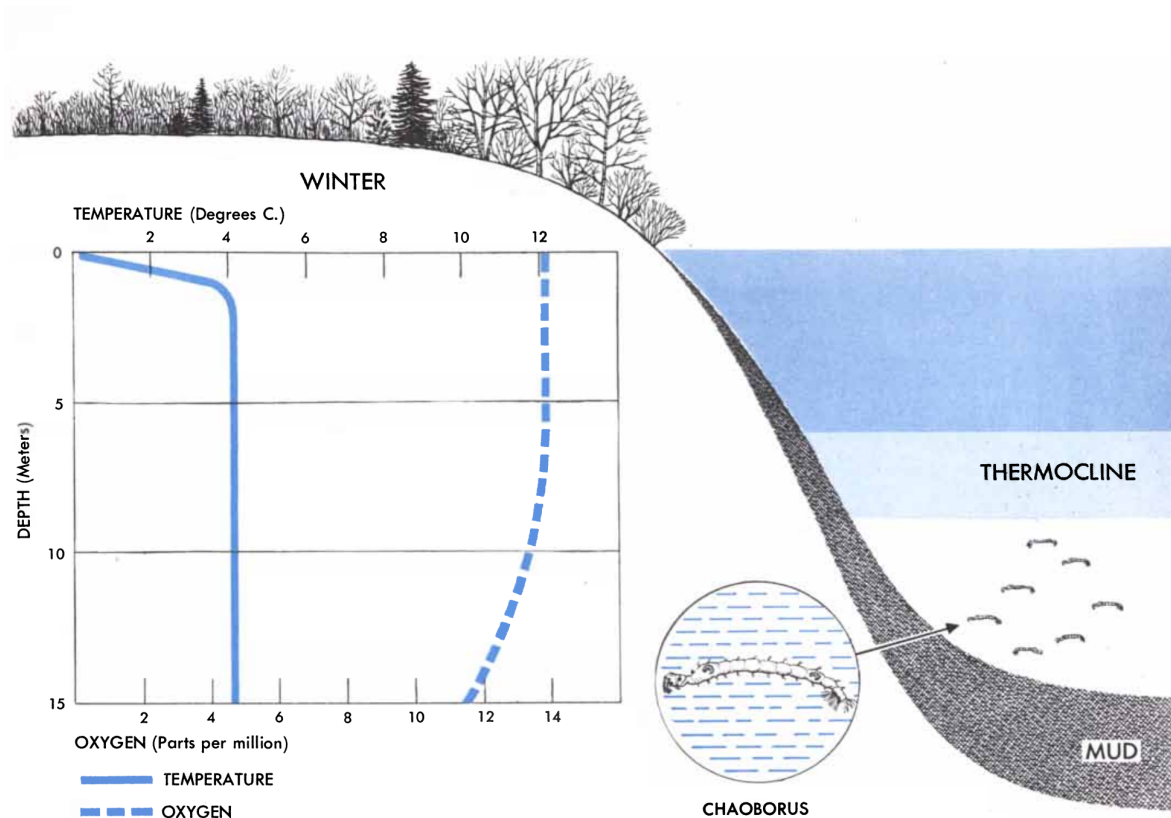
with that of the Earth System<sup>3,79,223,224,268,289,295</sup>. However, many levels of organization in this hierarchy seem to be populated by assemblages rather than systems. Does, for example, the population, more or less a collection of organisms from a single species, exhibit emergent features that make it a unity? The same doubt is apt for the community and the landscape level. Furthermore, this monolithic hierarchical depiction does not hold for all species, as multi-cellular organisms will have an additional cell layer, whereas single-celled organisms will not. Along the same lines, the hierarchy for social animals like ants or humans will, most probably, harbor additional social layers. Exceptions like these might be incorporated in the hierarchy model as *mesoforms* or by explicitly allowing branching hierarchies<sup>85,89,122,218</sup>. However, some authors have suggested that the hierarchy itself is a mere appearance arising from observations at different spatial and temporal scales<sup>2,104,209</sup>. Nevertheless, the concept of the levels of organization remains a helpful heuristic to structure scientific problems regarding life on earth and will act as such in this thesis<sup>36,46</sup>.

This conception of complex adaptive systems has some consequences for statistical analysis, to which I will return throughout the thesis. The first, and most important, is the role of the environment. Because it is defined as “everything but the system”, it contains an almost infinite number of parameters potentially relevant for the system of interest<sup>308,310</sup>. Because of this, it is not possible to control for all potential co-founders when performing correlation analyses. This obstacle for well-controlled correlations does not fully invalidate regression analyses but needs to be taken into account when interpreting their results<sup>38</sup>. Second, the interconnections between the parts of the system facilitate their communication; changes at one point of the system’s structure will propagate and, given the right conditions, affect the system as a whole<sup>20</sup>. Third, the system’s parts usually show a high degree of diversity, and this diversity is often constitutional for the system. Thus, there is a danger in aggregating parts or applying statistical methods (that work with average values, thus collapsing the diversity)<sup>184</sup> that is further aggravated by the impossibility of experimentally verifying the correctness or sufficiency of the operationalization. Fourth, the presence of emergence in systems suggests that non-linear behavior and relationships will be present in and between the system’s parts<sup>16,61</sup>. Of course, the inverse is not true: Non-linearity is not a sufficient indication of emergence as it can also be a product of multiple linear interactions in an assemblage<sup>25,329</sup>. Fifth and last, the hierarchical model reminds us that a process can be described in different ways when observed at different levels of organization. Therefore, counter-intuitive outcomes can be encountered when the level of observation is not made explicit in analysis<sup>4,182</sup>.

### 1.2.2 LAKE ECOSYSTEMS AS SYSTEMS

Let us now turn to the scene of the study at hand, the lake ecosystem. Broadly but inclusively defined, a lake is a water-filled basin formed by volcanic, tectonic, glacial, biological, or anthropogenic activity<sup>275,319</sup>. The water can stem from rivers, streams, melting ice, precipitate, or springs below the lake and is usually fresh or brackish water<sup>135</sup>. The geo-history and geomorphology of the lake’s basin characterize the lake so that the shape of the basin, as well as the lake’s water circulation characteristics, have been used to classify lakes<sup>144,149,224</sup>. In addition to the structure conferred by the lake’s basin, lake water is, in the summer, stratified into the epilimnion, the thermocline, and the hypolimnion. This stratification arises from warming of the surface of the lake (see figure 1.1).

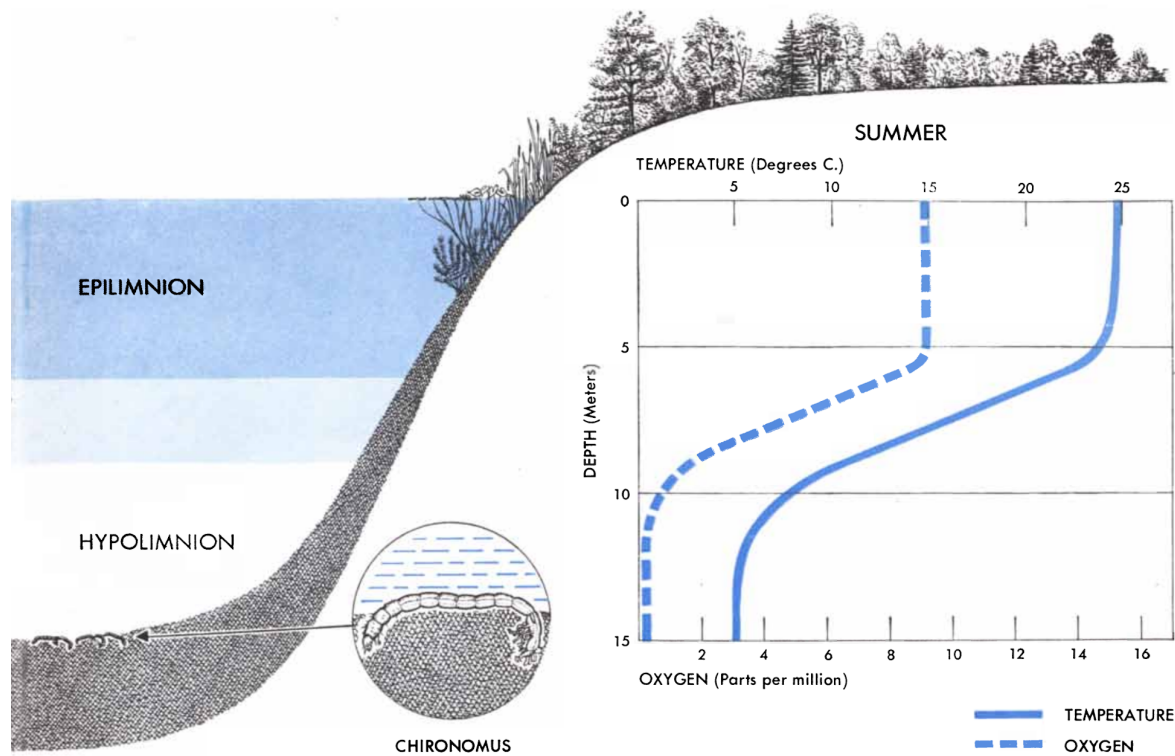
Lakes can be seen as islands of water surrounded by soil. This is the case because, in their re-



**Figure 1.1:** The stratified structure of a lake. In the summer, the upper part of the lake, the epilimnion, is warmed by the sun, leading to higher availability of oxygen in the water (inset on the right). The thermocline represents the rapid transition from the warm to the colder water of the hypolimnion, which also exhibits less circulation. In winter, lakes exhibit a rather uniform water temperature across the vertical axis (insert on the left). These differences constitute different niches for different organisms. This figure is taken from Deevey, *Scientific American* (1951) and depicts Linsley Pond with an exaggerated vertical scale<sup>67</sup>. (Continued in figure 1.2 on next page.)

spective landscape, lakes are rather isolated when compared to, for example, soil or forest habitats. Nevertheless, lakes are integrated into their geographical surrounding through their catchment<sup>127</sup>. The catchment is the area surrounding the lake from which a significant amount of precipitation will be transported to the lake's basin. A lake's water quality depends on the catchment's ability to purify water and decreases if the catchment is polluted<sup>229</sup>. With the water, substances ranging in size from atoms to small animals are transported into the lake, which acts as a (temporary) sink for these materials. This way, the lake reflects the the physico-chemical and ecological state of the its catchment. Because of this, a lake can be regarded a sentinel of environmental change of the landscape it is part of<sup>275,321</sup>.





**Figure 1.2:** The stratified structure of a lake. (Continued from figure 1.1 on previous page.)

Water has properties that make it essential for life, including its capability as a solvent, its large thermal capacity, and the water density anomaly<sup>301</sup>. Furthermore, most lakes feature a steady in- and outflow of water that constantly replenishes the nutrients present in the lake, which are dispersed throughout the lake through water. The stratification and the basin shape of lakes, in contrast, create habitats with distinct properties and, thus, different biotic dynamics and different compositions of the biotic communities<sup>23</sup>. This way, lakes can act as a habitat for many different organisms in different parts of the lake but also provide essential services for organisms that live on or around it<sup>301</sup>.

In contrast to the lake as a topographic entity, the lake ecosystem can be defined as the system that emerges from the interaction among biotic factors (i.e., the organisms living in the lake) and between biotic and abiotic factors (i.e., physical, chemical, geographical, etc.)<sup>295</sup>. In other words, the lake ecosystem can be thought of as the system in which the niches of the individual species

present in the lake overlap and interact<sup>153</sup>. Lake ecosystems show system-specific, i.e., non-linear, adaptive and emergent, properties<sup>119,128,178,221,222</sup>. For example, the lake ecosystem is not bounded by the lake itself, as many organisms live on or around the lake and still belong to the lake ecosystem. Furthermore, some processes in the lake's catchment exert an influence on the lake's biotic community that is so strong that they must be regarded as part of the lake ecosystem<sup>151,224,319</sup>. This leads to the practical problem that geographically close points in the lake's catchment might not be equally important for the ecosystem's functioning, leading to a spatially heterogeneous or structured catchment and a hard-to-define border for the lake ecosystem<sup>3,228</sup>.

Along similar lines, we can describe the relationship between the organisms in the lake ecosystem in terms of processes and interactions such as flows of energy or matter between individuals or species<sup>126,299</sup>. Decades of study have resulted in a categorization of species according to their position in the food chain: Primary producers or autotrophs acquire energy and material from sunlight and abiotic sources, respectively, and produce biomass “out of thin air”. These organisms serve as the fundament for a cascade of consumers, in which organisms at a certain level predate on the species at the next lower level of the chain. Decomposers, finally, break down detritus and make it available to the trophic cascade<sup>185</sup>. The organisms present in an ecosystem constitute its biotic community<sup>224,239,271</sup>. Through the interactions between species, the role of physical factors for a species in an ecosystem is co-determined by interactions with other species (a point that will be highly important for what follows)<sup>182,183</sup>. Nevertheless, the biotic community is only a part of the ecosystem and does not form a unit and must be considered an assemblage rather than a system.

### 1.2.3 ENVIRONMENTAL MICROBIOMES

Current estimates state that there are around  $10^{12}$  microbial species on Earth and that bacteria alone, despite their microscopic size, represent approximately 15% of all biomass on earth<sup>15,187</sup>. In all ecosystems, the majority of organisms are microscopic, regardless of their taxonomic placement. It is, therefore, not surprising that microbes are essential parts of all environmental ecosystems<sup>30,72</sup>. For example, prokaryotic and eukaryotic algae and some other protists act as primary producers while bacteria, archaea, and fungi decompose detritus, making them “ubiquitous janitors of the Earth”<sup>47</sup>. Furthermore, microorganisms are essential for the ecological cycling of nutrients such as nitrogen and phosphorus<sup>9,230,258,333</sup>.

In recent decades, the term “microbiome” has emerged. Since then, many definitions have been brought to the table that is synonymous in many but not all crucial aspects. Some authors derive the term “microbiome” from “biome”, thus including the microbes’ environment into the term; others derive it in broad analogy to the other *-omes*, i.e., as the total of all microbial organisms in an ecosystem. There have been attempts to standardize this nomenclature, but they leave some operational issues open<sup>24,197</sup>. The consensus that emerges from the literature is that the microbiome is equal to the microbial community composition, i.e., the microscopic part of the biotic community.

Following this consensus, the microbiome cannot be regarded a system as defined in section 1.2.1, because this would require some autopoietic processes that include all microbes in a given ecosystem but excludes, say, amphibians or insects. Such processes that separate by organism size are not known. The closest match are possibly emergent substructures in the microbiome like the microbial loop<sup>254,290</sup>. Instead of passing through the vertical food chain, ca 50 % of all primary production pass through the microbial loop that is essential for nutrient recycling and emerges from commensalism, competition, and predation between groups of microorganisms<sup>9</sup>. Nevertheless, we have to conclude that the microbiome is an assemblage with local substructures (or sub-systems)<sup>235</sup>, just like the biotic community that it is a part of.

As such, the microbiome is adapted and reactive to changes in the ecosystem it inhabits<sup>176,297,256</sup>. Recent studies show that the environment is the most significant factor in determining the microbial community composition<sup>11,112,198,252</sup>. Nevertheless, stochastic and historical processes such as uneven dispersal of microbes because of geographic barriers such as mountain ranges are far from insignificant for microbiome dynamics<sup>18,29,30,225,241</sup>. These insights were instrumental in falsifying the statement that “everything is everywhere, but the environment selects”, i.e., that microorganisms are neutrally dispersed (for example as “air plankton”) with subsequent environmental selection creating the differences in microbiomes in different habitats<sup>63</sup>.

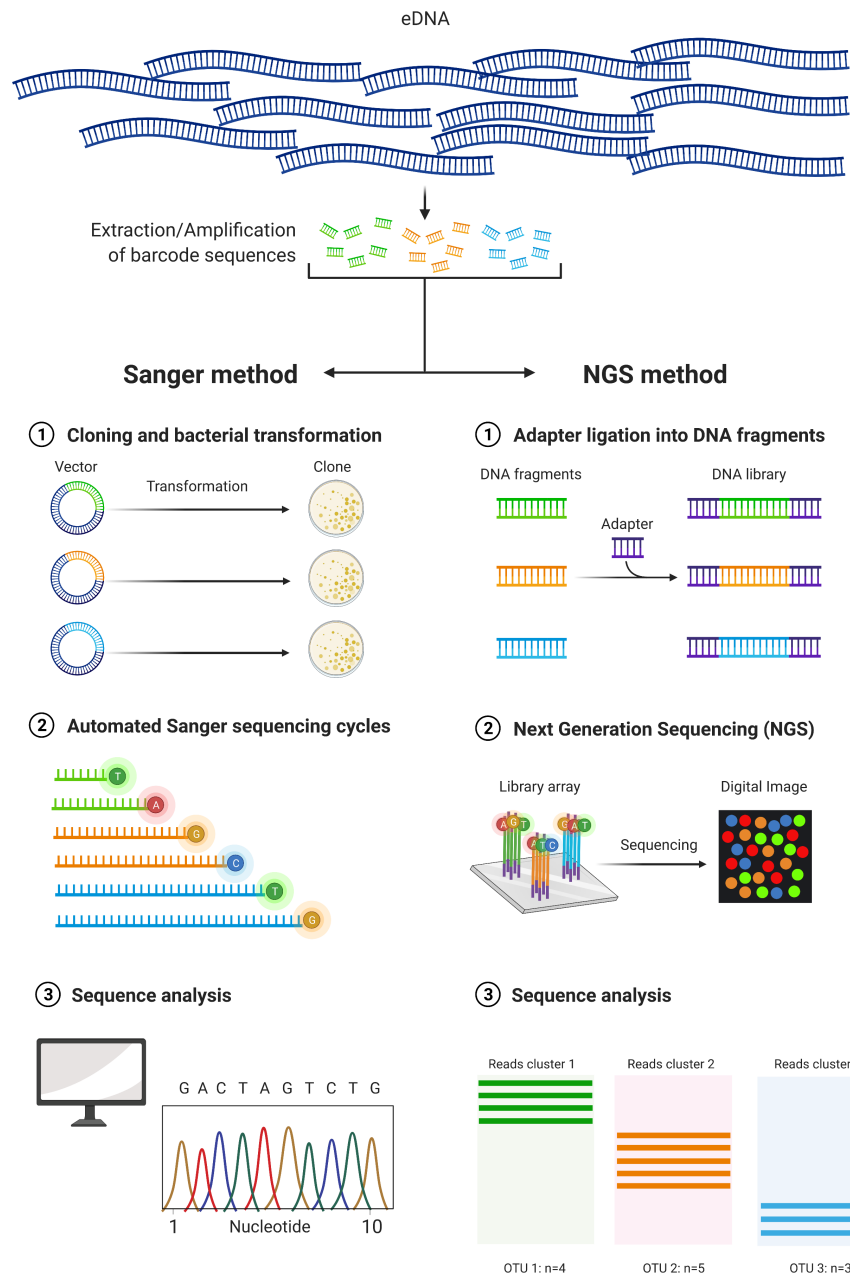
For this thesis, I want to use the thoroughly pragmatic definition of the microbiome as the assemblage created or defined by eDNA barcoding (for a description of this, see section 1.2.4). Arguably, this is the working definition for the microbiome in the literature since NGS methods have made it relatively easy to identify all – or at least, a good part of all – microbes present in a sample<sup>76</sup>. By explicitly adopting this methodological, pragmatic definition, I hope to guard myself and the reader against reification, i.e., confusion of the abstraction with the “real thing” or

an assemblage with a real unit<sup>148</sup>. This is necessary because the ease of using eDNA barcoding to determine the microbial community composition makes it easy to under-emphasize those aspects of the microbiome eDNA barcoding does not register. For example, all the interactions between microorganisms are lost in it, as well as the microdiversity below the taxonomic resolution of operational taxonomic units (OTUs)<sup>270</sup>. While the creation of OTUs can be seen as an aggregation necessary for further analysis (as discussed in section 1.3.2), microdiversity has been described as essential for ecosystem functioning in both theoretical and experimental studies<sup>103,178,210,296</sup>.

#### 1.2.4 SEQUENCING AND BARCODING OF eDNA

The ability to sequence the DNA of microbes is central to the study of microbiomes. This is, to a large extent, because only a small fraction of all microbes can, currently, be maintained in pure culture, and this has been a requirement for their study before the advent of *'omics* methods<sup>146,233,296</sup>. The first widely used DNA sequencing method was the chain-termination method, also known as Sanger sequencing. In its original form, it entails a modified polymerase chain reaction (PCR) with added dideoxyribonucleotide triphosphates (ddNTPs) acting as chain terminators and downstream identification of the fragment lengths using an agarose gel (see figure 1.3)<sup>174,269</sup>. After that, other methods for identifying DNA sequences in samples, such as enzymatic digestion or microarrays, were developed<sup>106</sup>. However, these methods suffer from drawbacks, including (i) poor scaling of runtime with the number of sequences (especially the case for Sanger sequencing) and (ii) the inability to identify previously unknown DNA sequences (i.e., unknown unknowns, especially the case for microarrays)<sup>284</sup>.

Both drawbacks are remediated in techniques collectively termed Next-Generation Sequencing (NGS) methods. Most importantly, NGS methods extend, modify, and adapt the Sanger sequencing method to be able to sequence multiple sequences at the same time, making these methods highly scalable<sup>284,293</sup>. For most NGS methods, this is achieved by integrating the amplification and detection steps of DNA sequencing by, e.g., using fluorescently tagged nucleotides, whose emission wavelength and intensity can be detected and used to reconstruct the sequence (see figure 1.3). The ability of NGS methods to increase the numbers of sequences generated in a single run by at least five orders of magnitude<sup>293</sup> enabled sequencing of whole genomes transcriptomes, i.e., the totality of RNAs present in a single cell or organism. Whereas for Sanger sequencing, it was often necessary to grow organisms extracted from environmental or patient-

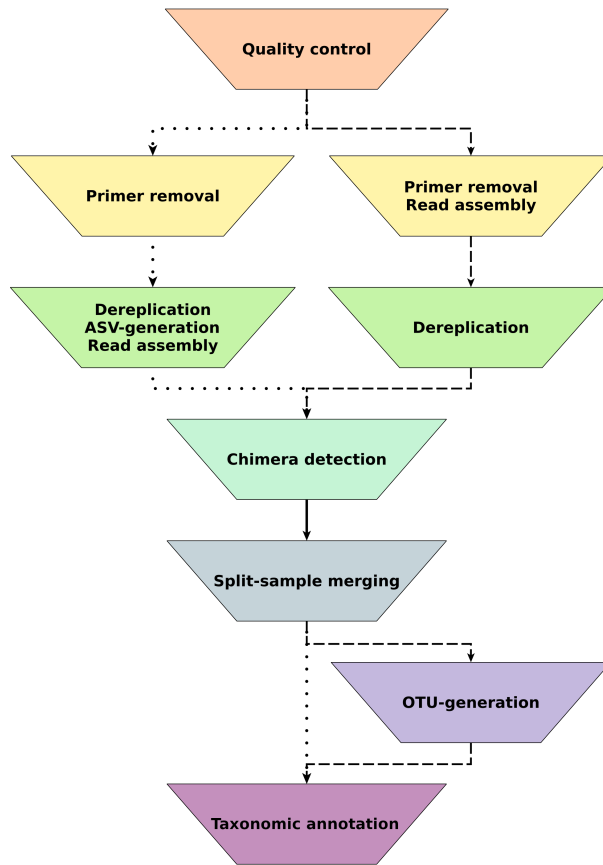


**Figure 1.3:** (Continued on the following page.)

**Figure 1.3:** (Previous page) DNA sequencing: Sanger vs. Illumina sequencing. After some preparatory steps to extract (i.e., excise or amplify) the region of interest from the total, purified eDNA in the sample, Sanger and Illumina sequencing take different steps. After creating a sufficient amount of DNA to sequence via, e.g., cloning (1, left), Sanger sequencing employs a modified PCR protocol with added chain terminators (2, left). Because the DNA polymerase can incorporate these in the place of one of the four nucleobases, the sequence can be read out by gel electrophoresis (given separate reactions for each of the nucleobases) or, given fluorescently labeled chain terminators, by capillary gel electrophoresis (3, left). On the other hand, Illumina sequencing requires the preparation of the DNA molecules by, e.g., ligation of DNA adapters (1, right). Then, the sequence of multiple DNA molecules can be read out in a single run because the elongation and base identification processes take place concomitantly (2, right). In contrast to Sanger sequencing, Illumina sequencing data requires more computational post-processing (see figure 1.4, but enables the counting of separate OTUs (3, right). Derived from “Shotgun Sequencing Sanger vs NGS”, by BioRender.com (2021). Retrieved from <https://app.biorender.com/biorender-templates>

related samples in pure cultures to increase the amount of DNA to sequence, NGS methods are able to sequence mixtures of DNA sequences<sup>294,307</sup>. Most relevant for this thesis, this includes environmental DNA (eDNA), i.e., all the DNA present in an environmental sample, encompassing DNA inside of microorganisms and viral particles, shed cells, excretions of macroscopic organisms, and extracellular DNA molecules<sup>238</sup>.

These new study objects bring novel insights: environmental metagenomics can hint towards which genes might be evolutionarily selected for, and environmental metatranscriptomics show which genes are actively transcribed in the studied environment<sup>298</sup>. However, when only interested in which species are present in a sample, it is more economical to focus the sequencing effort on so-called DNA barcodes, i.e., short DNA fragments used to identify an organism as member of a certain taxonomic unit<sup>293</sup>. To act as a DNA barcode, a gene or genetic locus needs to be present in all organisms that are to be studied, have an evolutionarily conserved function, a slow mutation rate, and a sequence size large enough to capture a sufficient degree of variation but small enough to be sequenced with the chosen method<sup>324,325,335</sup>. For metabarcoding, i.e., the identification of multiple organisms in environmental samples, the most frequently used genetic loci are on the small subunit (SSU) rRNA gene<sup>18</sup>. This gene is part of the transcription machinery, which makes it essential and, thus, limits its apparent mutation rate and ensures its presence in each organism. Because of differences in the ribosome architecture between Prokaryotes and Eukaryotes, the homologous genes are called 16S rDNA and 18S rDNA, respectively.



**Figure 1.4:** The Natrix pipeline as an example of NGS data post-processing from sequences to OTUs. After a quality control step that removes sequence reads that are probably erroneously sequenced, sequences are clustered into ASVs using DADA2 (left branch) or assembled and dereplicated into separate reads using pandaseq and CD-HIT (right branch), which involves the removal of sequencing primers from the sequences. Following that, OTUs are removed that might stem from chimerization of two distinct sequences in a prior amplification step. Because eDNA datasets are often sequenced bilaterally in split samples, it is, then, necessary to merge these two paired sequences before OTUs are generated and counted (only for those samples that were not already used to create ASVs in an earlier step). As the last step, the ASVs or OTUs are taxonomically annotated using widely used taxonomic databases, and all results are written to files. Figure taken from Welzel *et al.*, BMC Bioinformatics (2020)<sup>317</sup>.

Often, the eDNA barcoding a set of samples results in a table of operational taxonomic units (OTUs; OTU tables). After the sequencing itself, the application of a pipeline of software tools is necessary to, e.g., remove adapter sequences or technical artifacts and aggregate sequences with minor variations to OTUs (see figure 1.4). Different methods for sequence aggregation exist, but they roughly work as follows: A set of OTUs is defined based on all the sequences in the dataset. Then, the occurrence of each of the OTUs in each of the samples is counted. The classic method

and the one used here defines the OTUs by clustering SSU rRNA sequences that exhibit a sequence similarity of 97% or more<sup>196</sup>. OTUs defined this way are dependent on the structure of the dataset they originate from and are not comparable across datasets. A more recent method instead identifies amplicon sequence variants (ASVs) by clustering identical sequences after removing technical artifacts using a dedicated sequencing error model. OTUs generated this way are sample-independent and capture more variation than the former, but might lead to an overestimation of biodiversity present in the sample<sup>40,41,103,273</sup>.

Generally, OTUs generated with either of the two methods can be assigned regular taxonomic labels by looking up the OTU sequence in a dedicated SSU rDNA taxonomy database. In practice, however, this has limitations. For one, a very high percentage of the global microbial biodiversity remains unstudied, leading to massively incomplete databases<sup>68,296</sup>. Furthermore, because of the length of the DNA barcode and the aggregation of sequences into OTUs, we cannot resolve an OTU's taxonomy to a level lower than the Genus<sup>37,156</sup>. Finally, there is a tension between the phylogeny and the taxonomy of an OTU<sup>234</sup>. The former reflects the evolutionary divergence of a set of organisms. Assuming that the differences between the DNA barcodes of two organisms stem from rather uniform and piecewise mutation (instead of larger-scale changes such as insertions), the dissimilarity between their respective barcodes (as, e.g., given by the Levenstein distance) will correlate with their evolutionary distance<sup>324,325</sup>. As such, the phylogeny can be derived from the barcodes themselves. In contrast, taxonomy provides an extrinsic, labeled, and hierarchical classification of all organisms into levels of the taxonomy (with numerous exceptions stemming from the history of systematics, especially for Eukarya), ranging from the Domain, through the Phylum, Class, Order, Family, and Genus, and ending with either the species or the strain. However, the taxonomic classification of microbial as well as macrobial species is the source for much debate<sup>234,325</sup>. Taken together, potentially relevant variation and information can be lost when working only with taxonomically annotated OTUs instead of all the OTUs generated from a set of samples.



### 1.3 COMPUTATIONAL MODELS FOR MICROBIOMES

#### 1.3.1 GENERAL REMARKS

After the basic terms of this thesis are defined, the stage is now set to analyze the microbiome computationally. To this end, I will, in this section, present an overview of a few different methodological approaches used for computational analyses of ecosystems and biotic communities.

Note that the list of methods discussed here is not exhaustive but only contains methods relevant to this thesis's focus, which consists of studying environmental microbiomes.

I will use a uniform notation throughout the following subsections and will only deviate from this for variables and parameters unique to a given formula, method, or approach. Generally, matrices are referred to by bold, uppercase letters and vectors by bold, lowercase letters; regular, uppercase letters indicate cardinality, such as the number of species in a given sample. The dataset used in these examples is taken to be generated using eDNA barcoding and contains  $M$  samples  $\mathbf{m} = \{m_1, m_2, \dots, m_M\}$  taken from sufficiently comparable sites. In these samples, a total of  $N$  distinguishable OTUs are identified with  $\mathbf{n}_j = \{n_{1j}, n_{2j}, \dots, n_{Nj}\}$  designating the occurrences of all OTUs in sample  $j$ ; the number of occurrences of OTU  $i$  in sample  $j$  is given by  $n_{ij}$ . Furthermore, a set of  $E$  environmental parameters are measured for each of the samples. In analogy to the OTUs, the values of the environmental parameters measured for sample  $j$  are given by  $\mathbf{e}_j = \{e_{1j}, e_{2j}, \dots, e_{Ej}\}$  and the value of the environmental parameter  $g$  for sample  $j$  is given by  $e_{gj}$ . In a slight deviation from this nomenclature, let  $n_i$  stand for the number of OTU  $i$  in a sample if a single site is sampled repeatedly.

#### 1.3.2 NETWORKS MODELS

Because of the central role that interactions and processes play in the definition of systems in general (see section 1.2.1) and ecosystems more precisely (see section 1.2.2), network structures appear to be the logical representation of the microbiome. This is reflected by the fact that trophic chains and networks have been used as a description of ecological processes since, at least, the 18th century and maybe even the Medieval era, and have become even more popular in graphical form 1920s<sup>77,78</sup>.

In ecological networks (or trophic webs), each node represents a population or an aggregation of multiple populations, and the edges represent interactions or flows of energy or nutrients<sup>138,139</sup>.

Aggregation is necessary to reduce the number of interactions and flows one has to measure to a feasible level and theoretically permitted for sufficiently similar groups of organisms<sup>137,299</sup>. Commonly, aggregation is performed according to the taxonomy or the ecological function of the populations in question. Excessive aggregation, however, hides heterogeneity in the aggregates and can lead to ecological bias, i.e., misleading networks and models<sup>116,162,184,204,330</sup>. In the case of trophic networks, edges only represent trophic relationships such as predation. However, just as for the aggregation of populations, a less specific edge definition leads to more flexibility of the network model<sup>179</sup>. Flexibility is especially necessary when studying microbiomes, where (i) trophic relations cannot be observed at the scale necessary for network creation, and (ii) a broad range of interactions, such as metabolic cooperation, communication through secondary metabolites, and displacement through environmental toxins such as fermentation products, are of relevance.

As long as the effects of the interactions are quantifiable, they can be represented by the matrix  $\mathbf{I}$  with  $N$  columns and rows. When describing microbial interactions, the values  $\mathbf{I}_{i,j}$  give the numerical result of interaction, co-occurrence, or similarity between the populations or OTUs  $i$  and  $j$  on the occurrence of  $i$ . These can be derived from time-series observations of the species in question. In contrast, for microbiomes, the matrix  $\mathbf{I}$  is often inferred from OTU tables based on the assumption that strongly interacting OTUs will co-occur or have similar occurrence numbers in the same samples<sup>42,68,226</sup>. A wide range of methods have been developed for this, ranging from simple co-occurrence and correlation coefficients to methods that take the statistical properties of OTU tables, like compositionality and sparseness, into account<sup>141,159,245</sup>.

There is a wide range of approaches that extend the networks by a quantitative aspect. Early models were derived from economics and cybernetics, focused on the flow of energy between compartments, i.e., nodes with state variables, and were known as input-output models<sup>87,91,126,138,236,292</sup>. Focusing on changes in the occurrences of the populations, we can describe the changes in the occurrence of OTU  $i$  using the differential equation

$$\frac{dn_i}{dt} = f_i(n_1, n_2, \dots, n_N, e_1, e_2, \dots, e_E) = f_i(\mathbf{n}, \mathbf{e}), \quad (\text{I.1})$$

where  $f_i$  is an arbitrary function<sup>182</sup>. The simplest variant of this formula can be achieved under the assumption that the relationships between the organisms are linear and independent, leading

to

$$\frac{dn_i}{dt} = \sum_j^N \alpha_{ij} n_j + \sum_g^E \beta_{ig} e_g, \quad (1.2)$$

where  $\alpha_{ij}$  and  $\beta_{ig}$  are parameters representing the fitness of the OTU  $i$  accrued from interacting with one unit of  $n_j$  and  $e_g$ , respectively. Here,  $\alpha_{ij}$  are the values that populate **I**.

Of course, to create more realistic models, one could choose functions of higher degrees as well as non-linear functions in the place of  $f$  in equation 1.1, such as generalized Lotka-Volterra equations<sup>88,114,207</sup>. However, the central problem with this approach is its dimensionality. That is to say, the relatively simple linear model in equation 1.2 has  $N^2 + N \cdot E$  open parameters; more complicated would have an even higher dimensionality. In most cases, these parameters would need to be approximated based on measurements of the variables. To determine parameters in systems of differential equations, one needs at least as many samples as parameters are present. Therefore, this approach would necessitate a prohibitively high number of samples. Furthermore, many of the parameters might not be measurable, because of insufficient measuring methodology or technology, because of principal reasons (e.g., because the parameters are mathematical abstractions without material counterparts), or because the parameters are too vaguely defined<sup>179,181,271</sup>.

Based on the notion of complex adaptive systems as described in section 1.2.1, further criticisms against the methods described in this section can be formulated. Most importantly, neither the networks nor the models described here can account for levels of organization or emergence. While strict hierarchical relationships might be modeled using multi-level networks or hierarchical input-output models<sup>157,259</sup>, it is not clear how the organization of systems can be represented in networks or matrices. Furthermore, it is not clear how one would mathematically capture the relationships between operationally open emergent systems, their environment, and their parts (and their respective parts and environment and so on). Yet another problem arises from the fact that network structures derived from OTU tables are not guaranteed to capture the communities' interaction structure<sup>44</sup>. In principle, the methods employed for this measure a kind of similarity (such as Pearson correlation) between the counts of two OTUs over all samples, which does not guarantee (nor require) interaction. As indicated in section 1.2.1, complex environments are riddled with co-founders for correlation-based similarity measurements; this is particularly a problem for trivial inference methods like Pearson or Spearman correlation. We

must conclude that a complete model of microbiome dynamics will be impossible, at least with the current methodology<sup>109,179,183,220</sup>. That being said, the relative position of a node in a network still indicates its ecological function, and even partially specified models can lead to qualitative insights into ecosystem functioning<sup>181,182,292</sup>. Along similar lines, network structures inferred from OTU tables have been used to predict the ecological importance of groups of OTUs, although the biological relevance of this is debated<sup>13,14,257</sup>.

### 1.3.3 ECOLOGICAL STABILITY AND BIOMONITORING

While it might be impossible to model the total dynamics of the microbiome, we should still be able to make meaningful statements about how the microbiome responds to changes in its environment. For example, we might formulate this in terms of a concept of stability, i.e., the response of an ecosystem when faced with a disturbance. Ecosystem stability has attained considerable interest and has even been declared the most important question in ecology, but whether it can be solved in such a way as to generalize across different ecosystems and lead to practically implementable solutions is unclear<sup>117,124</sup>.

Yet, biomonitoring programs, such as under the Water Framework Directive (WFD) of the European Union, come close. These programs record changes in ecological integrity over time by repeatedly sampling the same ecosystem, habitat, or site. The current state of the ecosystem, as represented by the sample's physical, chemical, and biological make-up, is then compared to a reference state, i.e., a state of the ecosystem in question object to only low levels of anthropogenic stress<sup>132,168</sup>. Deviations from this “pristine” state are formalized as Ecological Quality Ratio (EQR), based on which ecosystem management decisions are devised to achieve or stabilize “good ecological status”<sup>31,134</sup>. Currently, more than 300 different monitoring schemes for aquatic ecosystems are in use throughout Europe, differing in the ecosystem type they monitor and in the way the ecosystem's state is registered<sup>26,133</sup>.

In theory, the best way to get a full image of the current status of an ecosystem is to measure all its parameters. Of course, in practice, this is impossible. However, as the parts of the ecosystem interact (which they must for the ecosystem to be a complex adaptive system), ecosystem parameters are not independent but reflect the state of other ecosystem parameters<sup>271,291</sup>. Therefore, the design of a monitoring scheme has to strike a balance between two adverse goals. On the one hand, one might want to maximize the number of ecosystem parameters measured because more

parameters allow for more insight into the state of the ecosystem as a whole ecosystem state. On the other hand, one would like to minimize the cost of measurement, which rises with the number of different measurement techniques used and, to a lesser degree, the number of parameters measured. Many monitoring schemes under the WFD overcome this impasse by using biotic indices, i.e., composite metrics based on the number and identity of a pre-defined set of organisms present in the ecosystem<sup>26,133,277</sup>. The organisms surveyed this way are bioindicators, i.e., species whose presence or abundance reflects the state of their environment<sup>201</sup>. Bioindicators are like the proverbial canary in the coal mine that alerted the miners to danger before mechanical early warning systems were developed<sup>253</sup>. Instead of a single canary, however, most biomonitoring schemes exploit multi-species assessments to be sensitive to different environmental signals<sup>92,173</sup>.

To act as bioindicators, species or taxa need to occur in a broad range of habitats, be highly reactive to changes in their environment, and straightforwardly distinguishable<sup>58</sup>. While the first two requirements are readily met by wide ranges of microbes<sup>253</sup>, only the advent of metabarcoding made it possible use not only diatoms and phytoplankton, but also morphologically cryptic species as bioindicators<sup>26,134,194</sup>. The inclusion of bacteria and archaea is probable to improve many biomonitoring schemes because these appear to be as potent sentinels for a wide range of ecological parameters, including pollution, eutrophication, and general ecosystem health<sup>29,125,178,278</sup>. Furthermore, in addition to recording the presence, absence or number of a high number of organisms in a single measurement run, metabarcoding can be more easily automated because it does not require the high degree of “hands on” expertise that is required by morphological identification. Taken together, these aspects have lead to arguments that eDNA metabarcoding should replace morphological identification of bioindicators in biomonitoring schemes<sup>49,50,51,101,134,307</sup>.

Potential bioindicators can be identified based on the indicator value function<sup>59,60,75</sup>. For this, the sites need to be clustered into distinct groups, e.g., according to the EQR gradient, and into three site groups designating good, average, and bad ecological integrity. The indicator value of species, OTU or taxon  $i$  and the site group  $j$  is the product of the specificity and fidelity of  $i$  and  $j$ , formalized as  $a_{ij}$  and  $b_{ij}$ , respectively. It can be calculated using

$$\text{INDVAL}_{ij} = a_{ij} \cdot b_{ij} = \frac{n_{ij}}{n_i} \cdot \frac{C_{ij}}{C_j} \quad (1.3)$$

where  $n_{ij}$  and  $n_i$  refer to the number of individuals of  $i$  present in the site group  $j$  and in gen-

eral, respectively, and  $C_j$  and  $C_{ij}$  refer to the number of sampling sites in cluster  $j$  in general and at which  $i$  was found, respectively.  $A$  reflects the predictive power of  $i$  for a site to belong to  $j$  (“specificity”) and  $B$  indicates how frequently  $i$  is found at sites belonging to  $j$  (“fidelity”) <sup>60</sup>. The value of  $\text{INDVAL}_{ij}$  will be high if and only if  $i$  is a good indicator of the site group  $j$  because it occurs at all sites belonging to  $j$  and only at these. A  $p$  value, representing the probability of the result based on the null hypothesis of no association between  $i$  and  $j$ , can be calculated from a permutation test <sup>60</sup>. More recent implementations of this method also calculate indicator values for species at combinations of groups (e.g., species that occur specifically at sites with either good or bad integrity, but not at medium integrity) to better reflect non-linear species-environment relationships <sup>60,162</sup>. Before a species is included in a biomonitoring scheme, the result of the indicator value analysis needs to be validated experimentally or with further observations <sup>163</sup>.

While we might want to criticize the theoretical underpinnings of biomonitoring – namely, the assumption that the reference state of an ecosystem is a kind of “natural state” and therefore preferable to any other state, as well as its disregard for potential nonlinearities in the development of the ecosystem <sup>151,264</sup> – its underlying concept are helpful for the analysis of microbiomes. In essence, bioindicator analysis focuses on the apparent response of organisms to a gradient of an environmental parameter while ignoring any other sources of complexity in the ecosystem. Furthermore, from a mathematical point of view, the indicator value method (see equation 1.3) is agnostic to the identity of the indicated, as long as it is quantifiable or can be used to classify sampling sites, which makes this method highly flexible. Finally, it is important to stress that the bioindicator-indicated relationship is not necessarily a direct or causal one <sup>173,281</sup>. Still, a sufficiently high number of bioindicators for a range of abiotic factors will nevertheless point towards the structure of the microbiome.

#### 1.3.4 MACHINE LEARNING MODELS

The move from a knowledge- to a data-driven paradigm initiated by ‘omics methods like metabarcoding necessitates algorithmic support and novel computational methods <sup>161</sup>. In turn, this shift might explain the rising popularity of machine learning methods for the study of environmental microbiomes <sup>101,175</sup>. Machine learning describes a rather broad class of computational methods from data science and statistics that perform classification or regression tasks <sup>64,66,143,150</sup>.

Machine learning methods are usually categorized as either supervised, unsupervised, or reinforcement learning, of which only the first category is relevant for this thesis. In formal terms, supervised learning is used to search a space of functions for a function that maps or projects a set of  $M$  input (or independent) variables or features  $\mathbf{X}$  to a single output (dependent or response) variable  $\mathbf{y}$  given a dataset  $\mathcal{D} = \{y_i, \mathbf{x}_i\}_{i=1}^N = \{y_i, x_{i1}, \dots, x_{iM}\}_{i=1}^N$ , with  $N$  samples (or instances)<sup>71</sup>. In other words, using supervised learning, one attempts to model the processes that shape the relationship between  $\mathbf{X}$  and  $\mathbf{y}$ . To this end, supervised learning models tune a set of model parameters to minimize an error function  $E(\mathbf{X}', \mathbf{y}')$ , where  $\mathbf{X}'$  and  $\mathbf{y}'$  are the training subsets of  $\mathbf{X}$  and  $\mathbf{y}$ , respectively. Additionally, many machine learning models implement model-specific hyper-parameters, such as the number of trees in a random forest or the penalty parameter  $C$  of a support vector machine. While hyper-parameters are not automatically tuned in model training, some machine learning packages perform hyper-parameter tuning by employing, e.g., grid search and elaborate training schemes.

Of the error functions available for regression models, the root mean squared error (RMSE) and the coefficient of determination  $R^2$  are most relevant for this work. Given the predicted values  $\hat{\mathbf{y}}$  for and the average value  $\bar{y}$  of  $\mathbf{y}$ , and the number of samples  $n$  in the (sub-)dataset, these metrics can be calculated using

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2}, \quad (1.4)$$

and

$$R^2 = \frac{\sum_i^n (y_i - \hat{y})^2}{\sum_i^n (y_i - \bar{y})^2}. \quad (1.5)$$

As can be seen from these formulae, the RMSE is zero if  $\mathbf{y} = \hat{\mathbf{y}}$ , has no upper bound and is of the same magnitude and unit as  $\mathbf{y}$ . In contrast, the  $R^2$  assumes a value between 0 and 1 and represents the fraction of the variation in  $\mathbf{y}$  that can be explained by the variation in  $\hat{\mathbf{y}}$  – and therefore, the amount of variation on  $\mathbf{y}$  that the model can explain.

The performance of a machine learning model is assessed using the same error metrics, but with regard to the testing dataset, i.e., a subset of the initial dataset different from the one used for training. The separation of data used in training and testing is necessary to ensure that the

model’s projection function has learned generalizes, i.e., not merely reproduces the training dataset but also performs well on other samples, additional data, or future observations. This is even further aggravated by high-dimensional, sparse  $\mathbf{X}$ , i.e., datasets with a high number of features and a low number of non-zero values<sup>73</sup>. To this end, both the training and testing sub-datasets need to be representative of the whole dataset, i.e., they need to contain samples that cover the variance of the real-world process one wants to approximate. In cases where the initial dataset does not have a high number of samples, there exists a pay-off: The larger the test set, the more probable it is that the ML model “sees” all relevant patterns in the training phase. At the same time, this also leads to a smaller and potentially biased test set, which might not contain all relevant settings of the full dataset. To combat this, it is possible to use cross-validation instead of a single data split: In it, the dataset is sliced into  $k$  near-equally sized sub-datasets, and a model is trained on  $k - 1$  slices and evaluated using the held-out slice. While for medium-to-small datasets, 10-fold cross-validation can be used, in which models are trained on 90% of the dataset and evaluated on the other 10%, very small datasets often require to leave-one-out cross-validation instead. In this training scheme, only one sample is used to evaluate the model trained on the other samples. In both schemes, we can evaluate the base model by handling the predicted values as if they were the predictions of a single model.

In the following sections, I will describe the most important classes of regression models used in this thesis concerning their usefulness for microbiome modeling. By laying out their mathematical background, I will point out to what degree the different classes of machine learning models can (i) model non-linear relationships, (ii) learn based on a dataset for which the assumption of variable independence does not hold, and (iii) learn from datasets that contain more input features than samples. Note that I will not describe either artificial neural networks (ANNs) or more complex model ensembles like multi-target learning using model chains since they do not play a role in the publications included herein. Nevertheless, refer to section 3.4 for a discussion of the use of these in modeling microbiomes. Note also that, unless otherwise noted, the following is based on state-of-the-art machine learning textbooks<sup>27,323</sup>.

#### 1.3.4.1 LINEAR REGRESSION MODELS

Linear machine learning models can be seen as straightforward applications of highly developed statistical concepts in the context of machine learning. As such, they are focused on interpretabil-



ity, statistical validity and conceptual simplicity. However, these favourable properties are counteracted by a high number of parameters that need to be tuned to train a linear model. A basic form of linear regression approximates the relationship between  $y_i$  and  $\mathbf{x}_i$  can be given by

$$\hat{y}_i = \sum_j^M (x_{ij} w_j) + b = \mathbf{x}_i^T \mathbf{w} + b \quad (1.6)$$

where  $\mathbf{w}$  stands for a vector of weights,  $b$  for the intercept, and  $\mathbf{x}_i^T$  denotes the transpose of  $\mathbf{x}_i$ , so that  $\mathbf{x}_i^T \mathbf{w}$  denotes an inner product. There is a wide range of methods for estimate the weight vector  $\mathbf{w}$ , one of which is called least squares and is given by

$$\min_w \sum_i^N (y_i - \hat{y}_i)^2 = \min_w \sum_i^N \epsilon_i^2, \quad (1.7)$$

where  $\epsilon_i$  stands for the prediction error for sample  $i$  and the algorithmic details of which are not of interest for now. The behaviour of this weight estimation method can be modulated by applying regularization, leading to lasso, ridge, and elastic net regression, which can be written as

$$\min_w \sum_i^N \epsilon_i^2 + \lambda_1 \sum_j^M |w_j|, \quad (1.8)$$

$$\min_w \sum_i^N \epsilon_i^2 + \lambda_2 \sum_j^M w_j^2, \quad (1.9)$$

and

$$\min_w \sum_i^N \epsilon_i^2 + \lambda_1 \sum_j^M |w_j| + \lambda_2 \sum_j^M w_j^2, \quad (1.10)$$

respectively, where  $\lambda_1$  and  $\lambda_2$  are hyper-parameters of the model. Note that equation 1.10 can be turned into equations 1.8 or 1.9 by setting either  $\lambda_1$  or  $\lambda_2$  to zero. Models using equation 1.7 are susceptible to outliers in the training dataset, i.e.,  $w_j$  will assume erroneous values if  $x_{ij} \gg \bar{x}_j$ , decreasing the generalization ability of the model. This is alleviated by the additional terms in equations 1.8 – 1.10. In addition, the second term in lasso regression (equation 1.8) forces weights to assume the value of zero, consequently acting as feature selection method (see section 1.3.4.4)

and thus further improving the model's ability to generalize.

The formulation of linear regression given in equation 1.6 makes a few implicit assumptions. Most important among these are the assumptions of (i) linear independence and (ii) homoscedasticity of the features, i.e. the features are neither linearly correlated and show the same variance across their range of values, as well as (iii) normal distribution of  $\mathbf{y}$ . In contrast, generalized linear models, such as the *glmnet* model used in paper II (see section 2.2), allow for a distribution of the output variable that is taken from the exponential family, thus not sharing the third assumption. Generalized linear models can be described as having a three-part architecture, consisting of

1. an approximation of the distribution of  $\mathbf{y}$  that belongs to a certain family of exponential distributions parametrized by  $\theta$  (related to the mean of the distribution) and  $\tau$  (related to the dispersion of the distribution),
2. a linear model as given by equation 1.6, and
3. a link function  $g$  that relates the mean of the distribution to the predictions of the model, i.e.,  $\mu = g(\hat{y})$ .<sup>211</sup>

Because of this, generalized linear models are able to model non-linear relationships between input and output variables while still remaining linear in the weights  $\mathbf{w}$ . The weights are usually estimated using elastic net regression (see equation 1.10). However, these changes also lead to an increased number of variables and hyper-variables for generalized linear models as compared to regular linear regression, and the assumptions of linear independence and homoscedasticity remain in place.

#### 1.3.4.2 SUPPORT VECTOR REGRESSION

When working with “small-N-large-M datasets”, i.e., datasets with more features than samples, one of the issues arising in machine learning are very large parameter spaces that have to be tuned. Support Vector Regression (SVR) models are able to greatly reduce the model dimensionality (as compared to linear models) by employing maximum margin learning and so-called support vectors<sup>73,304</sup>. At its core, an SVR model learns a function of the form

$$\mathbf{y} = \mathbf{w}^T \phi(\mathbf{X}) + \mathbf{b}, \quad (1.11)$$

which is similar to equation 1.6, but introduces the function  $\phi$ , which stands for a transformation of  $\mathbf{x}$  into a high-dimensional space, i.e.,  $\Phi : \mathcal{X} \rightarrow \mathcal{V}$ . In contrast to linear models, the weights of SVR models are not estimated using least squares or a modification of it, but by using

$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ \text{s.t.} \quad & \xi_i, \xi_i^* \geq 0 \\ & y_i - \mathbf{w}^T x_i \leq \epsilon + \xi_i^* \\ & \mathbf{w}^T x_i - y_i \leq \epsilon + \xi_i, \end{aligned} \tag{1.12}$$

where  $\|\mathbf{w}\|$  is the magnitude of the vector of parameters  $w$ ,  $\xi_i$ , and  $\xi_i^*$  are slack variables, and both  $\epsilon$  and  $C$  represent hyper-parameters. In the two-dimensional space, this equation describes a tube around the function given by equation 1.11, and the edges of the tube have a distance of  $\epsilon$  to the function itself. Because data points falling into the tube are not relevant for the function's definition,  $\epsilon$  acts as an error tolerance hyper-parameter. Data points inside the tube are ignored, because for these  $\xi_i = \xi_i^* = 0$ , whereas  $\xi_i$  and  $\xi_i^*$  penalize points falling over and under the tube's edges, respectively, with the distance to the tube's edge. The influence of this penalization on the estimation of the weights can be tuned using the regularization hyper-parameter  $C$ . Finally,  $\|\mathbf{w}\|$  in equation 1.12 forces the function in equation 1.11 to be as flat as possible given the above limitations. Taken together, equation 1.12 formulates a multi-objective loss function with a trade-off between the maximal flatness of the function itself and a minimal size of the surrounding tube, allowing only  $\xi_i + \xi_i^*$  data points outside of the tube.

Instead of storing the vector  $\mathbf{w}$ , the SVR model only keeps track of the support vectors, i.e., those  $N_{SV} x_i$  that are outside the tube given by  $\epsilon$ . It is possible to calculate  $\mathbf{w}$  from the support vectors using the formula

$$\mathbf{w} = \sum_{i=1}^{N_{SV}} (\alpha_i^* - \alpha_i) \phi(\mathbf{x}_i) \tag{1.13}$$

that can be derived from equation 1.12 by finding the Lagrangian<sup>8,74</sup>. In this equation,  $\alpha_i^*$  and  $\alpha_i$  are the Lagrange multipliers of the constraints in equation 1.12 involving  $\epsilon + \xi_i^*$  and  $\epsilon + \xi_i$ ,

respectively. Thus, equation 1.11 can be rewritten as

$$y_i = \sum_{i=1}^{N_{SV}} (\alpha_i^* - \alpha_i) \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_i) + b. \quad (1.14)$$

Finally, there is the problem of calculating the inner (or dot) product  $\phi(\mathbf{x})^T \phi(\mathbf{x}')$ , which can, for some functions  $\phi$ , be written as Kernel function

$$K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \phi(\mathbf{x})^T \phi(\mathbf{x}'). \quad (1.15)$$

and calculated as kernel matrix or a kernel map. This way, it is not necessary to transform the whole data set, but only to calculate the function  $K$  for the support vectors. For the *svmLinear* model used in papers I and II (see sections 2.1 and 2.2), the kernel function takes the form of

$$K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle \quad (1.16)$$

and for the *svmRadial* model used in papers I and II (see sections 2.1 and 2.2) it is

$$K(\mathbf{x}, \mathbf{x}') = \exp \left( -\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2} \right), \quad (1.17)$$

where  $\sigma$  is an additional hyper-parameter.

Taken together, SVR models differ from the aforementioned linear models by (i) the optimization function given in equation 1.12, which improves generalization by allowing for regression errors in the range of  $2\epsilon$ <sup>32</sup>, (ii) the model sparsity, resulting from being derived from support vectors as described in equation 1.13 instead of the whole weight vector and (iii) their ability to approximate non-linear functions when using non-linear kernels.

#### 1.3.4.3 DECISION TREES AND ENSEMBLES OF DECISION TREES

The third and final class of machine learning models described here are regression trees and ensembles thereof. The decision trees used here are formalized versions of their popular counterparts: From a “root” node emanate two or more directed edges that represent an exhaustive set of possible outcomes of a decision (most of the time “yes/no” or “over/under”). The edges lead

to new nodes that themselves act as the root for a new set of decisions. This repeats until a “leaf”, i.e., a terminal node without further edges, is encountered. In machine learning, decision trees are created by recursive partitioning: Starting at the root node of the tree, one of the features in the training dataset is chosen and the dataset is split into two (by samples) so as to maximize the value of a splitting function. A minimal representation of the split as well as the average value of the dataset is stored in the parent node. Then, each of the node’s children is assigned the sub-dataset that fulfils the decision encoded by the edge that leads to this node. The choosing-and-splitting-process is applied recursively until either (i) the subset of independent variables at a node is homogeneous, (ii) splitting would not increase the value of the splitting function, or (iii) a predefined depth of the tree is reached. Finally, the decision tree resulting from this training procedure can be pruned, i.e., distal nodes can be removed, which can increase the ability of the model to generalize. To predict a  $y_i$  when given a sample  $\mathbf{x}_i$ , one starts with the root node, traversing the decision tree while always taking the edge that is indicated by the splitting decision of the node given the data at hand. Finally, the value of the leaf node is assigned to  $y_i$ .

In addition to its graphical depiction, a decision tree can be represented as a set of rules derived from the splitting decisions. That is to say, each node contains a decision rule of the form “If [VALUE] of feature [FEATURE] is larger than [SPLITTING POINT], then [NEXT RULE or FINAL VALUE], else [NEXT RULE or FINAL VALUE]”. This linguistic representation improves the interpretability of decision trees.

One example algorithm for a decision tree is the Classification And Regression Tree (CART) as implemented in the model *rpart* and used in paper I (see section 2.1)<sup>34</sup>. CART models are trained node by node, by calculating the splitting metric for each possible (binary) split and for each feature, and then the split-feature-configuration with the maximal splitting metric is determined at the current node. This is repeated recursively for all daughter nodes. The most widely used criterion to evaluate a split at value  $s$  of the feature in question, resulting in the splitting of the (sub-)dataset  $\mathcal{D}$  into  $\mathcal{D}_L$  and  $\mathcal{D}_R$  is derived from the analysis of variance (ANOVA) and can be written as

$$R(s, \mathcal{D}) = r(\mathcal{D}_L) + r(\mathcal{D}_R). \quad (1.18)$$

with

$$r(t) = \frac{1}{N} \sum_{y_n \in \mathcal{D}} (y_n - \bar{y}_{\mathcal{D}})^2, \quad (1.19)$$

where  $\bar{y}_{\mathcal{D}}$  is the average of the output variables in  $\mathcal{D}$  and  $N$  is the number of samples in  $\mathcal{D}$ .

Usually, single decision trees do not generalize well; instead, they over-fit to specific cases, which makes them “weak learners”. It has been shown, however, that “weak learners” can be used to construct very potent ensemble models, as long as the single models show enough variation<sup>160,272</sup>. Multiple methods are available to construct ensembles of decision trees with sufficient variation. The first one of these is known as bootstrap aggregating (or bagging), in which each of the CART models in the ensemble is trained on a randomly sampled subset of the original training dataset. This leads to bagging trees, such as implemented in the *treebag* model used in paper I (see section 2.1). Along the same line, Random Forests, such as implemented in the *rf* model used in paper I–III (see sections 2.1 – 2.3) are created by training decision trees on a random subset of the original dataset, but here, the subset is created by sampling features instead of samples<sup>33</sup>. For both of these models, the prediction for a sample is performed by traversing all the trees in the ensemble in accord with the sample  $\mathbf{x}_i$  and averaging the predictions of the single trees.

In contrast to the methods that create variance through randomization, gradient boosting means to construct each additional tree in such a way as to give the largest improvement over a loss function. This is accomplished by fitting the new tree to the residuals of a loss function, leading to a total function of the form

$$\hat{y}_i = \sum_{k=1}^K f_k(\mathbf{x}_i), \quad (1.20)$$

where  $K$  is the number of CART trees in the ensemble and  $f_k$  refers to the functions they implement<sup>96</sup>. In addition, stochastic gradient boosting, as implemented in the *gbm* model and used in paper I and II (see sections 2.1 and 2.2), employs the bagging method for tree fitting<sup>97</sup>. Extreme gradient boosting, as implemented in the *xgbTree* model used in paper I and II (see sections 2.1 and 2.2), further improves the gradient boosting algorithm on an algorithmic level, allowing fitting of more decision trees in the same runtime and introducing sparsity-aware algorithms<sup>43</sup>. As indicated in equation 1.20, the prediction  $\hat{y}_i$  of a boosted tree ensemble models is given by the sum of the predicted values of the single trees given an input  $\mathbf{x}_i$ .

Ensembles of decision trees have been shown to be very potent machine learning models as they boast a generally comparatively low prediction error. Furthermore, these models have a

list of properties that make them highly suitable for application to microbial ecology<sup>65,66</sup>. First, ensembles of decision trees are straightforwardly interpreted. Second, decision trees and their ensembles approximate non-continuous functions, which suits the potentially non-linear relationships between the microbiome and its environment. Third, they implicitly take dependencies between variables into account by way of their structure as stacked decision rules. Fourth, because they favour splits in “effective” features, ensembles of decision trees are able to handle high-dimensional datasets; in a sense, they perform feature selection. Fifth, and following from this, the dimensionality of a decision tree model does not scale with the dimensionality of the feature space.

#### 1.3.4.4 FEATURE SELECTION

A central issue when applying machine learning models is their ability to generalize from the training dataset to test data or other unseen samples. A central factors for this is the dimensionality of the input dataset, i.e., its number of features. Features that neither play a role in nor can be used to describe the real-world process that links input and output data can be seen as noise. A machine learning model can misinterpret this noise as signal based on the training dataset, reducing its ability to generalize. This is especially the case for *omics* datasets that contain a high numbers of features and a far smaller number of samples. Reduction of the dimensionality of the input dataset can, thus, improve model performance, but also accelerate model training, and increase the model’s understandability<sup>130,261</sup>.

Feature selection (FS) methods pick a subset of the original features by modelling, using some measure or method, the usefulness of the features for model training. In contrast to this, dimensionality reduction techniques, such as principal component analysis, project the input dataset’s full feature space into a new, lower-dimensional space but neither distinguish noise from signal nor contribute to model interpretability. Three different classes of feature selection methods can be distinguished<sup>213</sup>. The first class, embedded FS, is implemented in the machine learning models themselves. The lasso and elastic net regularization methods described in section 1.3.4.1 and the decision trees described in section 1.3.4.3 are examples of this class of FS. A second class of FS methods are filtering methods. These are deployed “in front” of the machine learning model, i.e., before model training itself, and they filter the input variables without explicit recourse to the model. The third and last class are so-called wrapper methods; these have access to the model and

can, therefore, select features to directly maximize the model's performance. In most cases, this involves iteratively training models while subtracting (backward selection) or adding (forward selection) features to the models training dataset. Because this requires multiple, subsequent training phases, wrapper methods are prohibitively time- consuming when tasked with a very high initial number of features. Because of this, no wrapper methods have been used in the work presented here.

Different filter methods select different sets of features because the methods they implement make different assumptions of what might constitute a “useful” feature. The correctness of these assumptions is, in turn, strongly related with the concept that the machine learning models will learn from the dataset at hand. Thus, there is “no free lunch”, i.e., there exists no single, globally optimal FS method. This makes it necessary to either experimentally determine the best-fitting out of a set of preselected FS methods for the given task or to employ ensembles of FS methods (as, e.g., implemented in the EFS package<sup>212,213</sup>). In many cases, the most expedient FS method is to manually filter the features based on domain expertise. However, in many cases one applies machine learning exactly because it is not clear which of the features are relevant. This is even more the case in data-driven (as opposed to expertise-driven) research contexts (see section 1.1).

In the work presented in this thesis, aside from the FS methods implemented in the machine learning models themselves, I used three filter methods. Firstly, I employed EFS in paper I (see section 2.1), modified in order to be able to handle regression tasks. The choice of the FS methods for paper II and III (see sections 2.2 and 2.3) have been motivated by prior ecological research. The first of these is the fast, correlation-based filter (FCBF). In essence, this method identifies groups of features that correlate with each other and then selects the one feature out of every group that correlates most closely with the dependent variable. As such, it models the relative dependencies between the OTUs that constitute the independent variables and is, thus, a multivariate approach. Furthermore, when applied to OTU tables, FCBF can be interpreted in the context of correlation networks that were described in section 1.3.2. Along these lines, FCBF can be interpreted as identifying hubs in the correlation network and then picking those OTUs from these hubs that correlate most with the respective target variable. The second filter method used here is the indicator value method as described in section 1.3.3. To that end, for each target variable, three groups of samples were created that showed high, medium and low values of the variable (i.e., the samples were stratified by tertiles of the target variable). Then, only those OTUs



were selected for model training that appeared as indicative (i.e.,  $\alpha \leq 0.05$ ) for at least one of the sample groups in question in order to account for variability in the breadth of the organisms' niches<sup>60</sup>. In contrast to the FCBF, this filtering method does not take into account any possible dependencies between the OTUs in the dataset, and is thus univariate.

## 1.4 DETAILS OF THIS WORK

### 1.4.1 DATASETS

The research presented in this thesis centers around two datasets collected from lakes. The first, used in paper I (see section 2.1), comprises data generated in August 2007 from 32 lakes that form an alpine transect in Austria<sup>121,217</sup>. It consists of an OTU table derived from 16s and 18s rRNA barcoding, the sampling date, and the GPS positions of the sampling points. The second dataset stems from Europe-wide sampling carried out in August 2012 and is used in papers II and III (see sections 2.2 and 2.3). This dataset contains 16s rRNA and 18s rRNA barcoding data for 280 and 218 samples, respectively. In addition to the sampling date, altitude, and the GPS position of the sampling point, this dataset also contains a small set of physico-chemical environmental variables (pH, conductivity, and temperature) for all samples, as well as an additional set of 21 physico-chemical variables for a subset of the samples<sup>29,30,219</sup>. In both datasets, each lake was sampled once at a water depth between 0.2 and 0.8 m near the shoreline, i.e., in the planktonic part of the epilimnion.

### 1.4.2 ASSUMPTIONS OF THIS WORK

In this last subsection of this introduction, I need to introduce operative assumptions. Some of these assumptions are necessary because of the characteristics of the data acquisition process. In contrast, others reduce the complexity of the study object to a degree where calculations are possible, and the understandability of the results is warranted<sup>62</sup>. It is necessary to make these assumptions explicit because they, in part, represent contradictions to the theoretical basis presented to this point. In a sense, the assumptions listed here construct a useful abstraction of the study object of this thesis; as such, the assumptions act like counterfactual handles that enable scientific work on the complex study object<sup>250,251</sup>. Explicitly enumerating these assumptions further allows me to explicitly refer to them and the reader to decide whether they are warranted. I begin the enumeration with two assumptions that stem from details of the generation of the datasets used here, continue with four assumptions necessary for statistical analysis of the dataset, and end with three assumptions essential for interpreting the results presented in this thesis.

- I. THE SAMPLES ARE REPRESENTATIVE. Many factors contribute to variation when sampling from ecosystems, such as (i) variation in sampling site, (ii) internal dynamics of the

sampld system, (iii) exceptional events such as irregular weather, and (iv) technical variation in sample post-processing. To control these sources of noise, one would need to take multiple samples from the same position, from different points of the same lake, over multiple days (i.e., biological replicates), or re-analyze the same sample multiple times (technical replicates). Due to the high cost of environmental sampling and sequencing, the datasets analyzed here do not contain replicates. Therefore, I need to assume that the data at hand are representative of the lakes and their microbiomes and that these do not reflect exceptional states of the ecosystems sampled.

- II. **LAKES ARE COMPARABLE.** To be able to perform almost any statistical analysis, it is necessary to assume that the samples are comparable to a sufficient degree. In statistical terms, regression analyses assume that the information present in the dataset is randomly drawn with the same underlying probability distribution. This means that the processes that form the relationships between the variables in the dataset need to be sufficiently similar, if not the same, throughout all samples. Thus, the dataset at hand needs to represent “normal” lake ecosystem dynamics in the absence of a disturbance. The degree to which this assumption is wrong will show up in the results as noise, possibly overshadowing signals. One can remove this kind of noise from the dataset by removing outlier samples, i.e., those samples from the analyses that seem not to be comparable to the others. To do this thoroughly requires a strong grasp of the processes underlying the generation of the measured variables. Instead, in the papers presented here, outliers were detected and removed based on statistical distributions of the measured environmental parameters.
- III. **EVERYTHING IS EVERYWHERE, BUT THE ENVIRONMENT SELECTS.** The first part of this assumption can be seen as describing a comparable, normal state, and thus following from the prior one. The second excludes historical, evolutionary, or geographical effects from the analysis, focusing on the environmental effects, i.e., the effects that the physico-chemical composition of the environment has on the microbiome. This is also an implicit assumption of indicator value analysis (as described in section 1.3.3). Although this assumption is most certainly wrong (as described in section 1.2.3), working without it would require additional, currently inaccessible data to correct for the dynamics of the historical and geographic spread of microbial species in both recent and farther history.

- IV. MICROBIAL SPECIES ARE EQUIDISTANT. Simple tabular structures implicitly suggest that each of their columns is independent of the other columns (the same goes for rows). In the OTU tables, the OTUs are represented as equidistant categories, i.e., without any indication of evolutionary or ecological similarity. For analyses that require the features (and thus, in this case, the OTUs) to be statistically independent, this assumption is necessary (as described in section 1.3.4). However, equidistance between OTUs is inconsistent with, e.g., their evolutionary phylogeny. Note that recognizing this issue has led to the development of methods such as weighted UniFrac, that take the phylogenetic distance of barcodes into account when, e.g., calculating beta-diversity distances between samples<sup>190</sup>. Similar approaches that can be used with machine learning models do not exist yet.
- V. ZEROES IN OTU TABLES REPRESENT ABSENT OTUs. One can see the generation of OTU tables from environmental samples as sampling in a statistical sense. Along those lines, we can imagine a probability distribution for each OTU that determines the likelihood of encountering this OTU at a certain abundance when taking and post-processing the environmental sample. The final OTU table contains integers drawn from those probability distributions. Thus, two possible interpretations of OTUs sampled zero times exist: The absence might either result from an actual absence of the OTU from the sample (or lake) or from the OTU being sufficiently rare. This is made more problematic by unique mathematical properties of the number 0, which can cause irregularities in statistical or machine learning analyses.
- VI. EACH OTU IS SUFFICIENTLY HOMOGENEOUS. During barcoding and the subsequent sequence analysis, all properties of an OTU except for its sequence and occurrence number are lost, and OTUs are, therefore, represented as a homogeneous class of enumerable objects. However, more and more studies show that there is a high and biologically relevant degree of variation between individuals aggregated in an OTU<sup>103,210,255</sup>. Furthermore, such heterogeneity is expected based on evolutionary and ecological theory<sup>184</sup>. Nevertheless, the definition of OTUs can be analogous to the aggregation necessary for model creation (as described in section 1.3.2).
- VII. OTUs ARE A RELEVANT UNIT OF ECOLOGY. The aggregation of individuals into OTUs by way of data generation automatically turns these into the smallest biological unit available for analysis. While OTU tables do not allow for any classification of units orthogonal

to this aggregation, it is still possible to create larger aggregates along the line of statistical properties or taxonomic annotation of the OTUs. Note that this assumption does not suggest that the OTU is the only relevant unit of ecology (or evolution, for that matter), as is sometimes discussed. On the contrary, from the theoretical basis outlined in section 1.2.1 follows that there are multiple levels of non-reducible, effective “units of ecology”, including all levels depicted in table 1.1, even if using OTUs as the smallest unit for a given analysis.

#### VIII. ENVIRONMENTAL PARAMETERS ARE NEGLIGIBLY DETERMINED BY THE MICROBIOME.

Just as their environment shapes organisms, organisms change their environment by interacting with it. I assume that the actions of microbes have a negligible effect on the level of environmental parameters compared with other factors, such as the effect of macrobes, weather, or geological processes. Note that this assumption does not entail that microorganisms do not change their environment, but only that their effect on the measurements is negligible. An assumption like this would be necessary for causal analyses because it removes the potential for circular reasoning from the analysis. In contrast, this assumption facilitates the conceptual separation of biotic and abiotic factors in the lake for the analyses presented here. Otherwise, e.g., an abiotic factor completely determined by the microbial community would need to be included in the microbiome if we follow the definition of a system in section 1.2.1.

#### IX. NOT ALL RELEVANT ENVIRONMENTAL PARAMETERS ARE IN MY DATASET, BUT ALL THE RELEVANT OTUS ARE.

This assumption is composed of two separate statements. The latter can be seen as an extension of the first assumption and underlines the OTU tables’ representativeness. Of course, this statement can lead to an underevaluation of rare OTUs and their importance for processes in the lake, but, in concert with assumption II, it allows me to speak of “the lake microbiome”. The former statement follows from the impossibility of recording all relevant parameters in a complex environment (see section 1.2.1). From this follows that correlations will always be under-controlled with regard to confounders. In turn, this makes it necessary to reinterpret the correlation-based results as, e.g., covariation, because the controls I can employ will not be sufficient to establish direct correlations.



# 2

## Publications

### 2.1 PUBLICATION I

Theodor Sperlea, Stefan Fuser, Jens Boenigk, and Dominik Heider (2018). SEDE-GPS: socio-economic data enrichment based on GPS information. *BMC Bioinformatics*, 19(440)

**Contributions:** T. Sperlea and DH conceived of SEDE-GPS, T. Sperlea and SF designed SEDE-GPS. SF wrote an initial version of SEDE-GPS, which was further improved on by T. Sperlea. T. Sperlea performed all analyses, supervised SF and drafted the manuscript. JB provided the lake dataset and discussed the results. DH supervised the project and revised the manuscript. All authors read and approved the final manuscript.

\* \* \*

The fact that ecosystems are located in space and sampling always happens at a specific point in time has positive and negative aspects for ecology. One of the positive ones is that the sampling point in space and time serves as a meta-datum with which additional data can be acquired that may assist in describing the ecological processes one studies. The core contribution of this study is a tool called SEDE-GPS (derived from *socio-economic data enrichment based on GPS informa-*

tion that automatically acquires data for a given global positioning system (GPS) location and time stamp. More specifically, SEDE-GPS queries the following databases or web services:

- **OpenStreetMap** for both land use of and points of interest in a predefined, circular area around the sampling point (20 and 73 features, respectively)
- **Eurostat** for a wide range of socio-economic features for the NUTS (*Nomenclature des unités territoriales statistiques*) region of the sampling point (for a total of 17 523 features)
- **Climate Data Center** for the average weather of the day, the month, and the year the sample was taken (12 features)
- **Twitter** for the number of tweets geo-tagged to points in the predefined area around the sampling point in the last seven days (1 feature).

Because it is not helpful for most analyses to gather and use all the features mentioned above, SEDE-GPS allows users to specify the sources they want to query and the subsets of features they want to gather. To exemplify the use of SEDE-GPS and get a preliminary insight into socio-economic impacts on lake microbiomes, we trained machine learning models to predict the microbial alpha-diversity in a set of alpine lakes. Our results show that a high degree of predictiveness can be achieved using *xgbTree* and *svmLinear* models. After filtering by an ensemble of feature selection methods (EFS), we examined the lists of features most important for predicting four different alpha-diversity metrics for Prokaryotes and Eukaryotes, separately. While these do not tell a clear story, it is striking that forest-related areas are important features for prokaryote alpha-diversity and not for Eukaryotes, while the contrary is the case for economic features.

\* \* \*

Ökosysteme sind im Raum verortet und Beprobung findet immer zu einem bestimmten Zeitpunkt statt – das hat positive wie negative Auswirkungen auf die Erforschung von Ökosystemen. So dient der Ort und Zeitpunkt der Probenentnahme in Raum und Zeit als Metadatum, mit dem zusätzliche Informationen gewonnen werden können, die wiederum bei der Beschreibung der untersuchten ökologischen Prozesse helfen können. Der Beitrag dieser Studie ist ein Werkzeug namens SEDE-GPS (abgeleitet von *socio-economic data enrichment based on GPS information*), das automatisiert weitere Merkmale (*features*) eines gegebenen GPS-Standorts und Zeitstempels einholt, indem es die folgenden Datenbanken oder Webdienste abfragt:



- **OpenStreetMap** für Landnutzung als auch Points of Interest in einem vordefinierten, kreisförmigen Bereich um den Messpunkt (20 bzw. 73 Merkmale)
- **Eurostat** für eine Vielzahl von sozioökonomischen Merkmalen der NUTS-Region (*Nomenclature des unités territoriales statistiques*), in der die Probenahmestelle liegt (für insgesamt 17 523 Merkmale)
- **Climate Data Center** für das durchschnittliche Wetter des Tages, des Monats und des Jahres, in dem die Probe genommen wurde (12 Merkmale)
- **Twitter** für die Anzahl der Tweets, die in den letzten sieben Tagen zu Punkten im vordefinierten Bereich um den Probenahmepunkt geschrieben und geo-getaggt wurden (1 Merkmal).

Da es für die meisten Analysen nicht hilfreich ist, alle oben genannten Features zu sammeln und zu verwenden, erlaubt SEDE-GPS dem Benutzer, die Quellen, die er abfragen, und die Teilmengen der Features, die er sammeln möchte, festzulegen. Um die Verwendung von SEDE-GPS zu veranschaulichen und einen ersten Einblick in die sozioökonomischen Auswirkungen auf das Mikrobiom von Seen zu erhalten, haben wir Modelle für maschinelles Lernen trainiert, um die mikrobielle Alpha-Diversität in einer Reihe von alpinen Seen vorherzusagen. Unsere Ergebnisse zeigen, dass mit den Modellen *xgbTree* und *svmLinear* ein hoher Grad an Vorhersagekraft erreicht werden kann. Nach der Filterung durch ein Ensemble von Merkmalsauswahlmethoden (EFS) untersuchten wir die Listen der Merkmale, die für die Vorhersage von vier verschiedenen Biodiversitätsmetriken für Prokaryoten und Eukaryoten am wichtigsten sind. Obwohl diese nicht eindeutig zu deuten sind, ist es auffällig, dass bewaldete Gebiete wichtige Merkmale für die Alpha-Diversität von Prokaryoten sind, aber nicht für Eukaryoten, während das Gegenteil bei wirtschaftlichen Merkmalen der Fall ist.

SOFTWARE

Open Access



# SEDE-GPS: socio-economic data enrichment based on GPS information

Theodor Sperlea<sup>1</sup>, Stefan Füsler<sup>1</sup>, Jens Boenigk<sup>2</sup> and Dominik Heider<sup>1\*</sup>

From BBCC Conference 2017

Naples, Italy. 18 - 20 December 2017

## Abstract

**Background:** Microbes are essential components of all ecosystems because they drive many biochemical processes and act as primary producers. In freshwater ecosystems, the biodiversity in and the composition of microbial communities can be used as indicators for environmental quality. Recently, some environmental features have been identified that influence microbial ecosystems. However, the impact of human action on lake microbiomes is not well understood. This is, in part, due to the fact that environmental data is, albeit theoretically accessible, not easily available.

**Results:** In this work, we present SEDE-GPS, a tool that gathers data that are relevant to the environment of a user-provided GPS coordinate. To this end, it accesses a list of public and corporate databases and aggregates the information in a single file, which can be used for further analysis. To showcase the use of SEDE-GPS, we enriched a lake microbial ecology sequencing dataset with around 18,000 socio-economic, climate, and geographic features. The sources of SEDE-GPS are public databases such as Eurostat, the Climate Data Center, and OpenStreetMap, as well as corporate sources such as Twitter. Using machine learning and feature selection methods, we were able to identify features in the data provided by SEDE-GPS that can be used to predict lake microbiome alpha diversity.

**Conclusion:** The results presented in this study show that SEDE-GPS is a handy and easy-to-use tool for comprehensive data enrichment for studies of ecology and other processes that are affected by environmental features. Furthermore, we present lists of environmental, socio-economic, and climate features that are predictive for microbial biodiversity in lake ecosystems. These lists indicate that human action has a major impact on lake microbiomes. SEDE-GPS and its source code is available for download at <http://SEDE-GPS.heiderlab.de>

**Keywords:** GPS, Data enrichment, Database, Ecology, Microbial ecology

## Background

The global positioning system (GPS), established in 1972 and made publicly available in 2000, allows for the exact identification of every spot on the surface of the earth [1]. Consequentially, when studying geographically localized objects or processes such as ecosystems, their location can easily be specified using GPS coordinates.

Many natural processes are strongly influenced by characteristics of their surroundings, i.e., it is known that chemical composition, size of different habitats, and

socio-economic features such as human population size, can influence the (microbial) biodiversity in ecosystems [2–5]. Therefore, having access to environmental characteristics and including them in analyses is crucial when trying to understand natural processes.

In the current study, we describe the novel tool SEDE-GPS (Socio-economic data enrichment based on GPS information), which can be used to enrich data sets with data from public and publicly available corporate databases based on user-specified GPS information. The current version of SEDE-GPS accesses Open Street Map (OSM), the Climate Data Center (CDC), Eurostat, and Twitter. SEDE-GPS has an easy-to-use graphical user interface and enables researchers to enrich their data with environmental and socio-economic information based on

\*Correspondence: [dominik.heider@uni-marburg.de](mailto:dominik.heider@uni-marburg.de)

<sup>1</sup>Faculty of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, D-35032 Marburg (Lahn), Germany  
Full list of author information is available at the end of the article



GPS information. This may lead to new insights into the influence of environmental and socio-economic features on a wide range of processes.

As an exemplary use-case of SEDE-GPS, we use it in order to identify features that have an impact on microbial biodiversity. To this end, we calculate different alpha diversity metrics from a sequencing dataset sampled from a set of alpine lakes in Austria. We then use feature selection and machine learning methods to determine features from the output of SEDE-GPS that can be used to predict these alpha diversity metrics. Our results show that both microbial Eukaryotes and Prokaryotes are impacted by different environmental features. Nevertheless, for both domains, the area and number of city structures (or lack thereof) and other human-related features carry high predictive power.

### Implementation

SEDE-GPS can be used via both a graphical user interface (GUI) and a command line interface. As main input, SEDE-GPS takes a list of at least one GPS coordinate. Additionally, SEDE-GPS needs a set of parameters specifying which databases will be queried and restrictions on the subfields to be downloaded. In the GUI, these parameters can be selected via mouse-click, however, in the command line version, these parameters need to be specified in a config file. The output of the different modules implemented in SEDE-GPS is temporarily saved and removed after being merged to a final output file in the csv format. This is due to the fact that the output of SEDE-GPS can be too large for regular-sized memory.

In the following, we will discuss the sources for data enrichment currently used by SEDE-GPS (Fig. 1).

#### OSM: Land use statistics

Open Street Map (OSM) is a community-generated, worldwide map. It is used by SEDE-GPS to gather information on land-use of the area that surrounds a given GPS position [6]. An area with an user-defined perimeter is extracted from relevant map tiles of the OSM database. As OSM maps are represented in Mercator projection, SEDE-GPS compensates for latitudinal distortion. From this map excerpt, the relative amount of pixels covered by different map legend objects are calculated by thresholding for their respective colors. This will calculate the fraction of area around the user-provided GPS position that is covered by, e.g., forests, city structures, or bodies of water.

#### OSM: POIs

In addition to the map itself, OSM also hosts a database that contains the locations of specific points of interests (POIs), such as special buildings or touristically relevant objects [6]. This module queries the OSM API and counts

the number of the different POIs in a perimeter of user-defined size around the GPS coordinates. As the OSM API reacts to queries slowly, this module is the largest contributor to the runtime of SEDE-GPS. Therefore, for larger analyses, it is advisable to manually download the so-called planetfile from OSM and to use it as an additional input for SEDE-GPS.

#### Eurostat: detailed regional statistics

The Eurostat database contains highly detailed governmentally collected data from the EU and EFTA member states [7]. Its regional database provides statistics on economic and social composition of centrally defined NUTS (*Nomenclature des unités territoriales statistiques*) regions. This module first determines the NUTS region that corresponds to the user-specified GPS position by querying the Google Maps database for the GPS positions' postal code. With around 17,500 features, this module's output represents 99.4% of all features gathered by SEDE-GPS.

#### CDC: European climate data

Via the CDC, a ftp server maintained by the Deutscher Wetterdienst (DWD), it is possible to publicly and freely access European climate data that dates back to 2010 [8]. The data has an interpolated spatial resolution of 5 km and a chronological resolution of a day or a month. This module requires a date as additional input and calculates average values of, e.g., temperature or windiness for the specified day, month, and/or year.

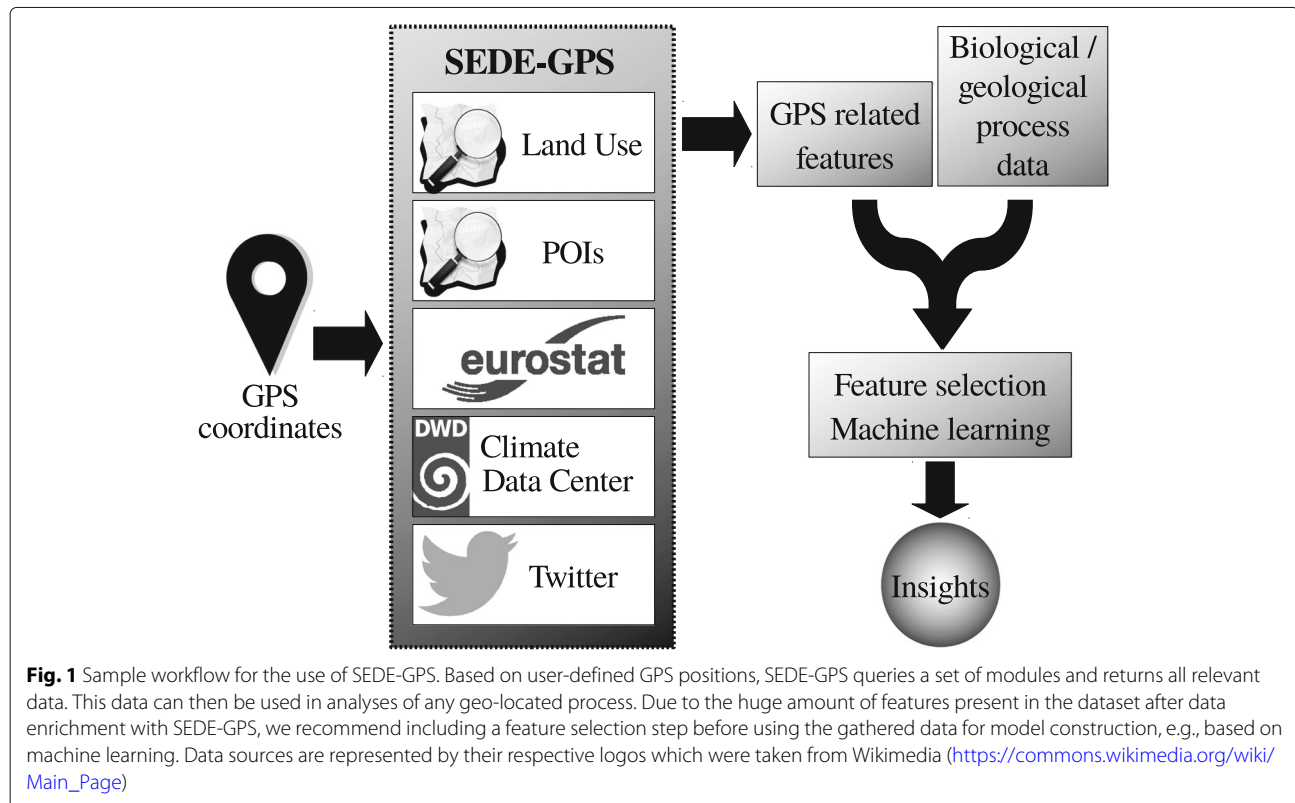
#### Twitter

The short messages sent out by users of Twitter (so-called tweets) can be location-tagged, and their number can be used to estimate tourist interest in a POI. The Twitter module of SEDE-GPS collects and counts tweets sent from a user-specified perimeter around the GPS coordinates. Twitter limits the access to its data so that SEDE-GPS can access all tweets that were sent in the last 7 days, but can only send 75 queries per 15 min. For a large number of GPS coordinates, this module will, therefore, require a long runtime.

## Methods

### Calculation of alpha diversity indices

The sequence data analyzed in the current study was taken from [9, 10] (Additional file 1). It stems from a set of alpine Austrian lakes, which were sampled in order to study the change of lake microbial ecosystems of three different lakes over time [9] and the difference in microbiome composition over many lakes [10]. 16s and 18s SSU rRNA sequences were sequenced using a 454 deep-sequencing amplicon approach [9, 10]. In the current study, only samples that were taken in August 2006 and contain more



than 1000 sequences were analyzed. 16s and 18s rRNA sequences were analyzed separately.

In order to estimate biodiversity within the samples, we calculated four different alpha diversity indices, namely Shannon's Entropy  $H'$ , Simpson diversity  $D$ , Simpson evenness  $E$ , and the Chao1 Estimator  $C$ , at the maximum possible sequencing depth with QIIME [11]. These indices describe the mean species richness or diversity at the local level [12] and are described by the following equations:

$$H' = - \sum_{i=1}^R p_i \ln p_i \quad \text{with} \quad p_i = \frac{n_i}{N} \quad (1)$$

$$D = 1 - \frac{\sum_{i=1}^R n_i(n_i - 1)}{n(n - 1)} \quad (2)$$

$$E = -\frac{1/\lambda}{R} \quad \text{with} \quad \lambda = \sum_{i=1}^R \left(\frac{n_i}{N}\right)^2 \quad (3)$$

$$C = R + \frac{S_1(S_1 - 1)}{2(S_2 + 1)} \quad (4)$$

where  $R$  is the number of species,  $n_i$  the number of individuals in species  $i$ ,  $N$  the total number of individuals,  $S_1$  the number of singletons (i.e., the number of species with

only one individual), and  $S_2$  the number of doubletons (i.e., the number of species with exactly two individuals).

#### Feature selection and feature evaluation

Before using the output of SEDE-GPS for machine learning, we employed a feature selection step. To this end, features containing missing values and with low variance (e.g., with more than 25% zeroes) were discarded. Next, we used the R package EFS (Ensemble Feature Selection) in order to rank the remaining features according to their importance. EFS is an ensemble learning feature selection method, that corrects for biases of the single methods when weighting features [13, 14]. Although EFS has been developed for feature selection in classification studies, we used an adapted version of EFS, which can be used for regression studies.

Stability of the features gathered over multiple runs of EFS were assessed by calculating the mean pairwise distance between the feature lists. To this end, we calculated Kendall's  $\tau$  and the Jaccard distance using the R packages *kendall* and *philentropy* [15, 16]. For two ranked lists of observations  $x$  and  $y$  of length  $n$ , Kendall's  $\tau$  is defined as

$$\tau(x, y) = \frac{c - d}{n(n - 1)/2} \quad (5)$$

with  $c$  being the number of pairs of concordant observations  $(x_i, y_i)$  and  $(x_j, y_j)$  with  $x_i < x_j$  and  $y_i < y_j$ ,  $d$  the number of discordant observations with

$$(x_i > x_j) \& (y_i < y_j) \parallel (x_i < x_j) \& (y_i > y_j), \quad (6)$$

$i$  and  $j$  indices in the lists  $x$  and  $y$ , respectively.

The Jaccard distance  $d_J$  for two lists  $x$  and  $y$  is defined as

$$d_J(x, y) = \frac{|x \cup y| - |x \cap y|}{|x \cup y|}. \quad (7)$$

Therefore, for two feature lists with a maximum distance, the Jaccard distance would assume a value of 1 and Kendall's  $\tau$  a value of  $-1$ . These values were calculated from feature lists that contain the 50 features that were ranked most important by EFS.

Sets of correlating features were determined using Spearman correlation at a correlation coefficient cutoff of larger than 0.7.

### Machine learning

We trained and evaluated eleven different machine learning models (as implemented in the R package *caret* [17]) using a leave-one-out cross-validation (LOOCV) scheme. These models included generalized linear models (*glmnet*), bayesian lasso (*blasso*), support vector machines (*svmLinear* and *svmRadial*), k-nearest neighbors (*knn*), Regression Trees (CART: *rpart*, bagged CART: *treebag*), Random Forests (*rf*), and stochastic and extreme gradient boosting (*gbm* and *xgbTree*). Models were evaluated by comparing the predicted values for all iterations to the real alpha diversity values, resulting in  $R^2$  values. Confidence intervals for the models' performance were calculated from the distribution of  $R^2$  values that were gathered from 1000x bootstrapped pairs of predicted and observed target variables. Their distributions were visualized using boxplots.

The machine learning models were tested for overfitting using a permutation test. To this end, the target variable was permuted and after feature selection with EFS, machine learning models were trained using the same approach as described above.  $R^2$  values were calculated and collected for 1000 repetitions of this procedure. Finally, the number of times  $t$  the resulting  $R^2$  value is larger than or equal to the  $R^2$  value received with an unpermuted target variable was counted. Significance in terms of a  $p$  value was calculated by  $p = t/1000$ .

## Results

### Data enrichment using SEDE-GPS

SEDE-GPS is structured modularly, with every module querying a certain database or API and, if necessary, data pre- and postprocessing steps (Table 1). The modules that query the Open Streetmap (OSM) databases,

e.g., have to account for the fact that their maps are in a Pseudo-Mercator projection or calculate a bounding box for counting of POIs. Some of the APIs queried by SEDE-GPS limit the number of queries that are handled in a certain amount of time (Twitter) or answer intentionally slowly (OSM). Similarly, the number of features provided by the different modules varies greatly, with Eurostat contributing by far the most the highest number of features, respectively (Table 1).

In order to showcase the use of SEDE-GPS, we planned to identify features that are predictive for the microbial biodiversity in a set of 39 alpine Austrian lakes. From these lakes, water samples were taken from which both 16s and 18s rRNA were sequenced and the geo-location of the sampling was recorded using GPS [9, 10]. These GPS coordinates were used as an input for SEDE-GPS, with all modules enabled, using radii of 1, 2, and 5 km and the date of sampling as additional input for modules for which this is necessary. This resulted in around 17,900 features.

The resulting dataset was observed to be highly sparse, with especially the output of the Eurostat and Twitter module showing a high degree of sparsity. Furthermore, a very small amount of features contained missing values, which we attributed to either errors in the databases or in the communication with the API. Therefore, features were discarded that contained any missing values or zeroes for more than a third of the instances. This procedure reduced the number of features per lake to around 1,200.

### Calculation of biodiversity metrics

The 16s and 18s rRNA sequencing datasets were processed separately using a QIIME pipeline [11]. Samples that contained less than 1000 sequences were discarded, which lead to differing numbers of lakes for which Eukaryotic and Prokaryotic biodiversity data were available. As biodiversity indicators, four different Alpha diversity metrics (Shannon's entropy, Simpson diversity, Simpson evenness, and the Chao1 estimator) were calculated after rarefaction ("Methods" section). We used multiple different metrics as they each measure biodiversity in specific ways and therefore emphasize different species distribution characteristics [18–20]. As the alpha diversity metrics were calculated for 16s and 18s rRNA separately, this resulted in maximally eight different biodiversity indicators for each lakes.

### Identification of important features using EFS

In order to find features in the output of SEDE-GPS that are predictive for lake microbial biodiversity, we used the R package EFS (Ensemble Feature Selection) and the eight alpha diversity metrics as target variable in separate analyses [13, 14]. EFS is an ensemble feature selection

**Table 1** Modules and their subfields currently available in SEDE-GPS

Module	Subfields	Additional Input	Data Processing	No. of features	Runtime (ms)
OSM Land Use	-	Radius	Pixel decompression	20	24823 ±2421
OSM POIs	Craft	Radius	Bounding boxes	7	3229 ±342
	Leisure	Radius	Bounding boxes	15	7202 ±622
	Powerplants	Radius	Bounding boxes	11	5053 ±503
	Special buildings	Radius	Bounding boxes	13	6881 ±453
	Tourism	Radius	Bounding boxes	8	3096 ±382
	Transport	Radius	Bounding boxes	13	6951 ±496
	Urban	Radius	Bounding boxes	6	2402 ±401
CDC	Average of the day	Date		4	<1
	Average of the month	Date		4	2 ±0
	Average of the year	Date		4	211 ±0
Eurostat	Agriculture			721	711 ±80
	Business Demography			778	1467 ±83
	Crime Statistics			4	16 ±4
	Demography			15077	2611 ±79
	Economic Accounts			67	431 ±41
	Education Stat.			30	31 ±5
	Labour Market Stat.			99	172 ±17
	Science & Technology			644	3718 ±400
	Tourism Stat.			44	163 ±11
Twitter	Transport			59	13383 ±224
	-	Radius		1	1014 ±316
Total				17629	83567

Runtime means and standard deviation were calculated from ten measurements

method that assigns weights to the features in an unbiased manner according to their predictiveness for the target value.

Using the average weight of the features as cutoff, features below this cutoff were discarded. To verify that the selected features are both descriptive and were not selected due to overfitting, eleven different machine learning models were trained to predict the eight alpha diversity values from the EFS-selected SEDE-GPS features. The models showed profoundly differences in performance (Table 2) with *xgbTree* showing near perfect performance for all target variables (Fig. 2). In order to confirm that the performance of the models is not due to overfitting, we performed a permutation test for the four best-performing machine learning models. For all target variables and machine learning models, this resulted in a p-value of less than 0.001.

Taken together, these results show that the features selected by EFS were not selected due to overfitting but are helpful for predicting alpha diversity metrics for prokaryotes and microbial eukaryotes in lakes.

### Stability and importance of features

Due to the fact that leave-one-out cross validation (LOOCV) was used to train and validate the machine learning models, multiple weighted feature lists were calculated for every target variable. Overfitting of EFS would have resulted in drastically different feature weights in the LOOCV iterations. In order to show that EFS did not overfit in the analyses presented here, we assess the stability of the features selected in the LOOCV iterations using both Kendall's  $\tau$  and Jaccard distance as feature list distance measures. These results show that the features selected by EFS show a high degree of stability and that the feature selection is not the result of overfitting (Fig. 3).

When manually examining selected features, it is important to keep in mind that the first step of feature selection in EFS is correlation based. This means that from sets of features that correlate, only the most descriptive feature is kept in the feature set. Therefore, for datasets processed with EFS, each feature label must be viewed as stand-in for a set of correlating features. Table 3 shows the five most important features for predicting the different alpha diversity metrics, with each feature name being



**Table 2** Performance ( $R^2$  values) of machine learning models trained to predict alpha diversity from SEDE-GPS output

Dataset	<i>glmnet</i>	<i>blasso</i>	<i>svmRadial</i>	<i>svmLinear</i>	<i>knn</i>	<i>rpart</i>	<i>treebag</i>	<i>rf</i>	<i>gbm</i>	<i>xgbTree</i>
Euk Chao1	0.292	0.003	0.713	0.980	0.0415	0.214	0.631	0.518	0.496	0.999
Euk Shannon	0.228	0.0167	0.791	0.993	0.000	0.180	0.635	0.582	0.680	1.000
Euk Simpson_e	0.277	0.0146	0.556	0.976	0.107	0.238	0.671	0.559	0.546	0.980
Euk Simpson	0.150	0.001	0.742	0.906	0.014	0.090	0.545	0.346	0.432	0.995
Prok Chao1	0.768	0.461	0.832	0.991	0.0695	0.420	0.635	0.915	0.955	0.979
Prok Shannon	0.527	0.011	0.940	0.991	0.172	0.538	0.626	0.930	0.993	0.999
Prok Simpson_e	0.345	0.128	0.849	0.991	0.035	0.304	0.622	0.937	0.840	0.999
Prok Simpson	0.459	0.008	0.915	0.986	0.168	0.453	0.627	0.904	0.880	0.991

replaced by higher order descriptions of the respective set of correlating features (for the simple feature names, see Additional file 2: Table S1). This examination was limited to five features per target variable because both the average feature weight and the stability of the feature position decrease quickly with increasing rank of the feature (Fig. 4, Additional file 3: Figure S1).

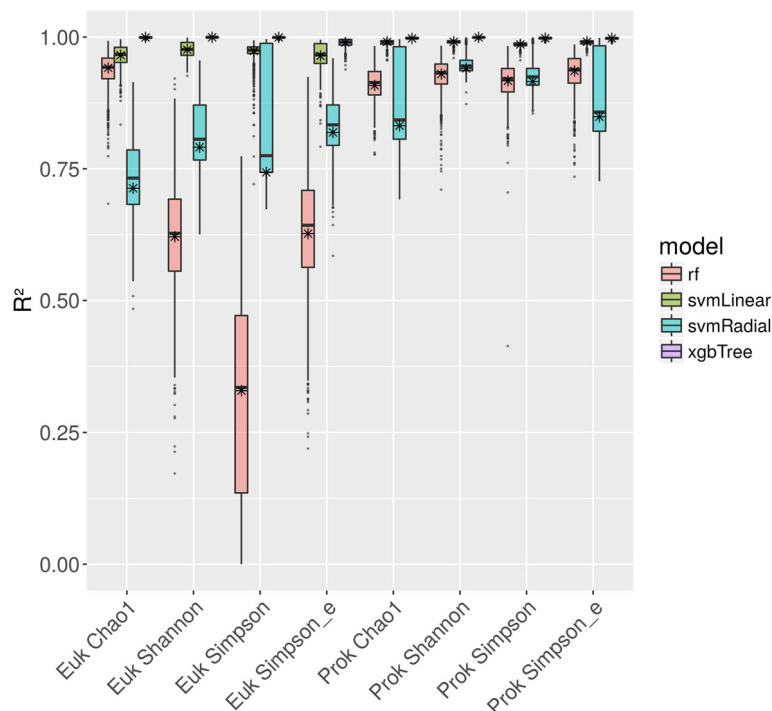
The resulting feature lists for Prokaryotes and microbial Eukaryotes show major differences, while using different alpha diversity metrics result, especially for Prokaryotes, in similar feature lists (Table 3).

## Discussion

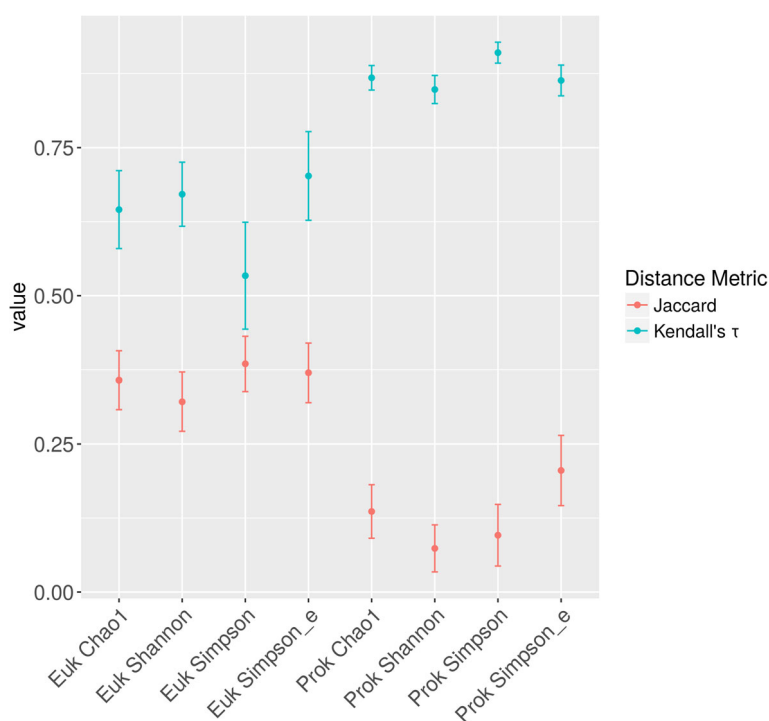
### SEDE-GPS

In this paper, we present SEDE-GPS, which can be used to drastically increase the number of features for datasets that contain GPS-located samples. Accessing four different data sources via five modules, it provides around 18,000 numerical features that contain socio-economic, geographic, and climate information (Table 1).

Currently, due to the choice of databases SEDE-GPS queries, this tool has a number of limitations. Both the CDC and Eurostat modules return only data for



**Fig. 2** Performance of machine learning models predicting microbial lake alpha diversity based on the output of SEDE-GPS. Stars represent the performance of models trained on the respective dataset, box plots represent confidence intervals of  $R^2$  values gathered from the respective model. Models were trained on the output of SEDE-GPS after feature selection and evaluated using LOOCV ("Methods" section). Only results for the four best-performing models are shown; for the others, see Table 2



**Fig. 3** Stability of feature lists over LOOCV iterations. Jaccard distances and Kendall's  $\tau$  were calculated for pairs of feature lists for the 50 most important features of each dataset. Dots and error bars represent average values and standard deviations of values, respectively. At maximum distance, the Jaccard distance and Kendall's  $\tau$  would assume a value of 1 and  $-1$ , respectively. Both feature lists are rather stable, however, the feature lists of the Prokaryote datasets are more stable than their Eukaryote counterparts

GPS coordinates in Europe, while the OSM modules and Twitter module will work for any GPS coordinate. Similarly, the databases queried by SEDE-GPS do not contain meaningful data for most marine GPS coordinates. In the future, we seek to overcome these limitations by including more data sources and thus extending SEDE-GPS both to new regions and to new data types and formats.

Similarly, the specific limitations and peculiarities of the databases currently used by SEDE-GPS are important for the interpretation of their data. OSM contains user-generated and user-curated information which might be of inconsistent albeit high quality or level of detail [6]. Eurostat, as a governmentally curated database, on the other hand, exhibits a level of detail which is generally lower than that of OSM as it can only be queried for defined NUTS regions [7]. As these regions are of widely differing sizes one might want to normalize data gathered from Eurostat to the area of the respective NUTS region. We decided not to implement this normalization step in SEDE-GPS as postprocessing steps not accessible to the user generally might introduce unwanted artifacts. The information gathered from Twitter comes with multiple caveats: For one, only very few processes will be directly influenced by the number of messages sent via

Twitter and this number will thus, in most cases, function as a proxy for other information. Additionally, the number of tweets will show a certain amount of variance over time, with the amount of variance being possibly also location-dependent.

Because of a rate limitation in API queries, both the OSM modules and the Twitter module are the biggest contributors to SEDE-GPS's runtime, especially for datasets with many GPS coordinates. It would be possible to speed up the OSM modules by reading the data from a so-called planetfile (an image of the OSM databases) instead of using API queries. This is, currently, not implemented in SEDE-GPS, as the planetfile is very large and a speed improvement would, therefore, only exist for very large GPS datasets.

Central to the design of SEDE-GPS is the fact that it does not perform any field-specific data postprocessing. Therefore, the output of SEDE-GPS can be used for studies in a wide variety of scientific fields. Nevertheless, for some applications, postprocessing steps might be advisable.

### Microbial ecology

In this study, we showcase the use of SEDE-GPS for microbial ecology. From the output of SEDE-GPS and using



**Table 3** Features with the highest weights for prediction of different alpha diversity metrics for Prokaryotes and Eukaryotes in Austrian lakes

Prokaryotes			
Chao1	Shannon Entropy	Simpson Diversity	Simpson Evenness
Industrial Area, Villages, Street (2-5 km)	Forests (5km)	Forests (5km)	Forests (5km)
Forests (5km)	Main street (small), married people	Forests	Main street (small), married people
Climate, Demography, City Structures	Forests (2km)	Buildings, Highways, Water, Parking, Parks	Forests (1km)
Climate, Demography, City Structures	Climate, Demography, City Structures	Forests (1km)	Buildings, Highways, Water, Parking, Parks
Main street (small), married people	Green space, small villages, Industrial area	Mining, main streets	Mining, main streets
Eukaryotes			
Chao1	Shannon Entropy	Simpson Diversity	Simpson Evenness
Forests	Main streets	Main streets	Economy (parking, GDP, Agrarian structures), Population
Family Demography	Beach & Water	Beach & Water	Economy (parking, GDP, Agrarian structures), Population
Climate, Demography, City Structures	Picnic Site (5km)	Economy (parking, GDP, Agrarian structures), Population	Beach & Water
Altitude, Climate, Demography, City Structures	Highway Pull-ins	Towns	Towns
Climate, Demography, City Structures	Urban regions, Av. Temperature, Parks	Urban regions, Av. Temp., Parks	Highway Pull-ins

For features in bold, a linear regression shows a positive relationship with the respective target variable

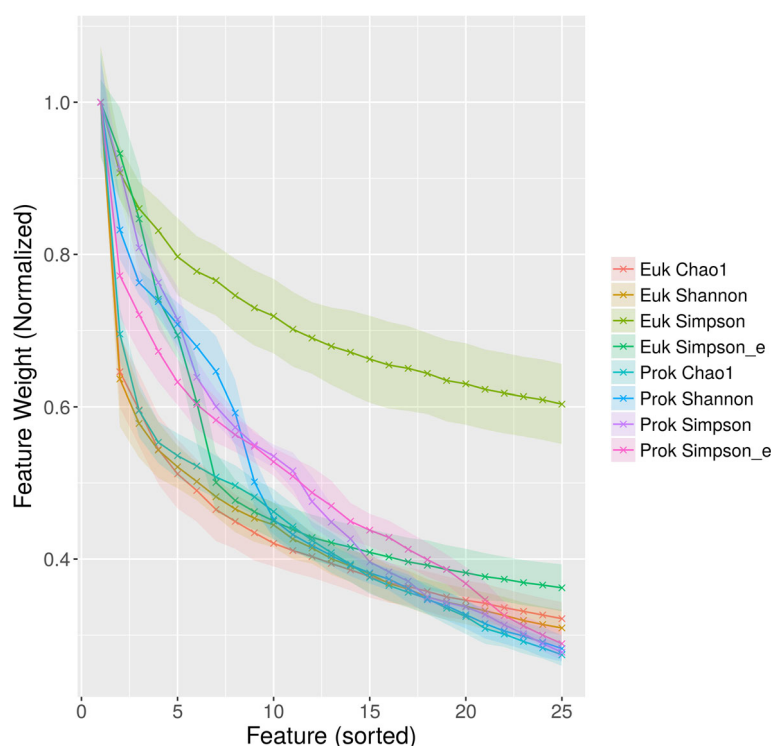
machine learning methods, we were able to identify features that can be used as predictors of both Eukaryote and Prokaryote alpha diversity in a set of alpine lakes.

Implicitly, in this study, we assumed that environmental features have a bigger impact on microbial biodiversity than historical contingencies and recent events. We acknowledge that this notion, succinctly formulated as “everything is everywhere, but the environment selects”, is highly debated [21–24]. Furthermore, we do not take into account that the composition of microbial communities can be majorly influenced by recent events or the microenvironment of the sampling position [25, 26]. These assumptions are necessary because the dataset analyzed here does not contain multiple samples that were collected on different time points for each of the lakes. However, we are not aware of such an ecological

microbial sequencing dataset with a quality, geographic extensiveness, and also uniformity of sample preparation comparable to the one we analyzed here.

The features we identified as most predictive for microbial biodiversity differed greatly between Eukaryotes and Prokaryotes, supporting the notion that microorganisms from these domains have different ecological roles [21, 24, 27, 28]. In contrast to this, the most predictive features for the different alpha diversity indices calculated from Prokaryotic sequences show a high degree of similarity. This indicates that the alpha diversity metrics used in this study essentially capture the same central distribution characteristics of the composition, at least for this domain of life.

Recently, many studies identified environmental and geographic features such as temperature, pH, climate, ion



**Fig. 4** Decline of average importance of features over the 25 highest ranked features. Feature weights were calculated using EFS and averaged over the LOOCV iterations. Ribbons indicate standard deviation. Average importance values were normalized so that the first feature has an average weight of 1. For all datasets except Euk Simpson, after the twelfth highest weighted features, feature weights are below 0.5

and nutrient concentration, and elevation-related environmental parameters as major drivers of the composition of lake microbiomes [4, 10, 21, 29–31]. Some of these features were also identified as highly impactful in our analysis (Table 3), albeit somewhat hidden under feature labels such as “Climate, Demography, City Structures” for temperature or “Economy (parking, GDP, Agrarian structures), Population” for nutrient concentration. While this clearly is a consequence of the field-agnostic nature of the data provided by SEDE-GPS, it might also point to possible sources for impact on biodiversity.

Therefore, our results also suggest that human action has a direct or indirect impact on lake microbiome composition. Although an impact of urbanization on biodiversity is well known for other areas of ecology [32–35], this is the first time, to our knowledge, that it has been described for microorganisms. Surprisingly, our results suggest that urbanization has a positive effect on Prokaryote biodiversity, as, e.g., the area of the environment covered by streets correlates positively with all biodiversity indices used in this study (Table 3). The negative impact of forest area might therefore stem from the fact that areas covered with forests cannot also be urban regions. Importantly, one should not fall into the trap of assuming that a higher biodiversity necessarily signifies a well-functioning

ecosystem [20] and take the results presented here to mean that more streets would improve lake ecosystems. Nevertheless, these results indicate that the processes that govern microbial ecology are very different from those that regard the ecology of larger organisms [9, 21, 28].

Further analyses will be needed to solidify the results of this study. In part, this is due to the fact that the samples and lakes included in this analysis are limited in number and are geographically close to each other [22, 24, 25, 36]. Therefore, for a more thorough analysis, larger datasets from more variable sites would be necessary, as currently only available from large-scale environmental sequencing efforts such as the Earth Microbiome Project [37] or the 1000 Springs Project [28, 38]. Nevertheless, on the basis of the results presented here, experiments can be designed in order to illuminate the mechanistic and causal relationships between environmental features and microbial biodiversity.

## Conclusion

This study shows how to use SEDE-GPS in order to enhance datasets that contain scarce amounts information on the environment of geo-located, observed processes. Analysing the output of SEDE-GPS leads to the identification of environmental, socio-economical, and

climate features that influence the studied process. These results can then act as basis for further hypothesis-driven research projects. SEDE-GPS is available at <http://www.SEDE-GPS.heiderlab.de>.

## Availability and Requirements

**Project name:** SEDE-GPS

**Project home page:** <http://www.SEDE-GPS.heiderlab.de>

**Operating system(s):** Platform independent

**Programming language:** Java

**License:** GNU GPLv3

**Any restrictions to use by non-academics:** None

## Additional files

**Additional file 1:** This table contains names, positions, and references for the samples contained in the sequence dataset and whether Prokaryotes and/or Eukaryotes were analyzed from the sample in this study. (CSV 3 kb)

**Additional file 2:** This table contains the feature names of the ten most important features in respect to the different alpha diversity metrics for Prokaryotes and Eukaryotes. Here, feature names were not replaced as described in "Methods" section. (CSV 2 kb)

**Additional file 3:** This figure shows the relative frequency of the most frequent feature at a given position for all target variables. Frequencies were calculated from the feature lists sorted by the weights determined by EFS in the LOOCV iterations. This shows that feature lists get more random with increasing rank of the feature on a sorted feature list. (TIF 844 kb)

## Abbreviations

CDC: Climate data center; GPS: Global positioning system; Lat: Latitude; Lon: Longitude; LOOCV: Leave-one-out cross validation; OSM: Open street map; POI: Point of interest

## Acknowledgements

Calculations on the MaRC2 high performance computer of the University of Marburg were conducted for this research. We would like to thank René. Sitt of HPC-Hessen, funded by the State Ministry of Higher Education, Research and the Arts, for installation and maintenance of software on the MaRC2 high performance computer. We would like to thank Julia Nuy for helping with data availability. OSM data copyrighted OpenStreetMap contributors and available from <https://www.openstreetmap.org>.

## Funding

This work was partially funded by the Philipps-University of Marburg. Publication costs for this manuscript were sponsored by the Philipps-University of Marburg.

## Availability of data and materials

Raw sequencing data can be accessed at the BioProject database under accession numbers PRJNA384345 and PRJNA384347.

## About this supplement

This article has been published as part of *BMC Bioinformatics Volume 19 Supplement 15, 2018: Proceedings of the 12th International BBCC conference*. The full contents of the supplement are available online at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-15>.

## Author's contributions

TS, SF, and DH conceived of and designed SEDE-GPS. SF wrote SEDE-GPS. TS performed the data analysis, supervised SF and drafted the manuscript. JB provided the lake dataset and discussed the results. DH supervised the project and revised the manuscript. All authors read and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>Faculty of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, D-35032 Marburg (Lahn), Germany. <sup>2</sup>Biodiversity Department, Center for Water and Environmental Research, University of Duisburg-Essen, D-45141 Essen, Germany.

Published: 30 November 2018

## References

- Parkinson B, Spilker J, Elkaim G. Global Positioning System (GPS). In: Mark H, editor. Encyclopedia of Space Science and Technology; 2003. <https://doi.org/10.1002/0471263869.sst069>.
- Vitousek PM, Mooney HA, Lubchenco J, Melillo JM. Human domination of earth's ecosystems. *Science*. 1997;277(5325):494–9. <https://doi.org/10.1126/science.277.5325.494>. <http://arxiv.org/abs/http://science.sciencemag.org/content/277/5325/494.full.pdf>.
- Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*. 2006;22(20):2532–8. <https://doi.org/10.1093/bioinformatics/btl417>.
- Hering D, Borja A, Carstensen J, Carvalho L, Elliott M, Feld CK, Heiskanen A-S, Johnson RK, Moe J, Pont D. The european water framework directive at the age of 10: A critical review of the achievements with recommendations for the future. *Sci Total Environ*. 2010;408(19):4007–19. <https://doi.org/10.1016/j.scitotenv.2010.05.031>.
- Grossmann L, Beisser D, Bock C, Chatzinotas A, Jensen M, Preisfeld A, Psenner R, Rahmann S, Wodniok S, Boenigk J. Trade-off between taxon diversity and functional diversity in european lake ecosystems. *Mol Ecol*. 2016;25(23):5876–88. <https://doi.org/10.1111/mec.13878>.
- OpenStreetMap contributors. Planet dump retrieved from <https://planet.osm.org>. 2017. <https://www.openstreetmap.org>. Accessed 15 Jan 2018.
- Eurostat. Eurostat database. 2017. <http://ec.europa.eu/eurostat/data/database>. Accessed: 21 Dec 2017.
- Deutscher Wetterdienst. Climate Data Center hosted by Deutscher Wetterdienst. 2017. <ftp://ftp-cdc.dwd.de/pub/CDC/>. Accessed: 21 Dec 2017.
- Nolte V, Pandey RV, Jost S, Medinger R, Ottenwälder B, Boenigk J, Schlötterer C. Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol Ecol*. 2010;19(14):2908–15. <https://doi.org/10.1111/j.1365-294x.2010.04669.x>.
- Grossmann L, Jensen M, Pandey RV, Jost S, Bass D, Psenner R, Boenigk J. Molecular investigation of protistan diversity along an elevation transect of alpine lakes. *Aquat Microb Ecol*. 2016;78(1):25–37. <https://doi.org/10.3354/ame01798>.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JL, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335–6. <https://doi.org/10.1038/nmeth.f.303>.
- Whittaker RH. Vegetation of the siskiyou mountains, oregon and california. *Ecol Monogr*. 1960;30(3):279–338. <https://doi.org/10.2307/1943563>.
- Neumann U, Riemenschneider M, Sowa J-P, Baars T, Kalsch J, Canbay A, Heider D. Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Min*. 2016;9(1):1. <https://doi.org/10.1186/s13040-016-0114-4>.
- Neumann U, Genze N, Heider D. EFS: an ensemble feature selection tool implemented as R-package and web-application. *BioData Min*. 2017;10(1):1. <https://doi.org/10.1186/s13040-017-0142-8>.

15. McLeod AI. Kendall: Kendall Rank Correlation and Mann-Kendall Trend Test. 2011. <https://CRAN.R-project.org/package=Kendall>. Accessed 15 Jan 2018. R package version 2.2.
16. Drost H-G. Philentropy: Similarity and Distance Quantification Between Probability Functions. 2017. <https://CRAN.R-project.org/package=philentropy>. Accessed 15 Jan 2018. R package version 0.0.3.
17. Kuhn M. Building predictive models in R Using the caret package. *J Stat Softw.* 2008;28(5):. <https://doi.org/10.18637/jss.v028.i05>.
18. Hill TCJ, Walsh KA, Harris JA, Moffett BF. Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology.* 2003;43(1):1–11. <https://doi.org/10.1111/j.1574-6941.2003.tb01040.x>.
19. Morris EK, Caruso T, Buscot F, Fischer M, Hancock C, Maier TS, Meiners T, Müller C, Obermaier E, Prati D, Socher SA, Sonnemann I, Wäschke N, Wubet T, Wurst S, Rillig MC. Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories. *Ecol Evol.* 2014;4(18):3514–24. <https://doi.org/10.1002/ece3.1155>.
20. Shade A. Diversity is the question not the answer. *The ISME J.* 2016;11(1): 1–6. <https://doi.org/10.1038/ismej.2016.118>.
21. Massana R, Logares R. Eukaryotic versus prokaryotic marine picoplankton ecology. *Environ Microbiol.* 2012;15(5):1254–61. <https://doi.org/10.1111/1462-2920.12043>.
22. Grossmann L, Jensen M, Heider D, Jost S, Glücksman E, Hartikainen H, Mahamdallie S, Gardner M, Hoffmann D, Bass D, Boenigk J. Protistan community analysis: key findings of a large-scale molecular sampling. *The ISME J.* 2016;10(9):2269–79. <https://doi.org/10.1038/ismej.2016.10>.
23. Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. The evolution of the host microbiome as an ecosystem on a leash. *Nature.* 2017;548(7665): 43–51. <https://doi.org/10.1038/nature23292>.
24. Boenigk J, Wodniok S, Bock C, Beisser D, Hempel C, Grossmann L, Lange A, Jensen M. Geographic distance and mountain ranges structure freshwater protist communities on a european scale. *Metabarcoding and Metagenomics.* 2018;2:21519. <https://doi.org/10.3897/mbmg.2.21519>.
25. Yi Z, Berney C, Hartikainen H, Mahamdallie S, Gardner M, Boenigk J, Cavalier-Smith T, Bass D. High throughput sequencing of microbial eukaryotes in lake baikal reveals ecologically differentiated communities and novel evolutionary radiations. *FEMS Microbiol Ecol.* 2017. <https://doi.org/10.1093/femsec/fix073>.
26. Jani K, Dhotre D, Bandal J, Shouche Y, Suryavanshi M, Rale V, Sharma A. World's largest mass bathing event influences the bacterial communities of godavari holy river of india. *Microb Ecol.* 2018. <https://doi.org/10.1007/s00248-018-1169-1>.
27. Guimarães PR, Pires MM, Jordano P, Bascompte J, Thompson JN. Indirect effects drive coevolution in mutualistic networks. *Nature.* 2017;550(7677):511–14. <https://doi.org/10.1038/nature24273>.
28. Oliverio AM, Power JF, Washburne A, Cary SC, Stott MB, Fierer N. The ecology and diversity of microbial eukaryotes in geothermal springs. *The ISME J.* 2018. <https://doi.org/10.1038/s41396-018-0104-2>.
29. Rossum TV, Peabody MA, Uyaguari-Diaz MI, Cronin KI, Chan M, Slobodan JR, Nesbitt MJ, Suttle CA, Hsiao WWL, Tang PKC, Prystajec NA, Brinkman FSL. Year-long metagenomic study of river microbiomes across land use and water quality. *Front Microbiol.* 2015;6:. <https://doi.org/10.3389/fmicb.2015.01405>.
30. Zeglin LH. Stream microbial diversity in response to environmental changes: review and synthesis of existing research. *Front Microbiol.* 2015;6:. <https://doi.org/10.3389/fmicb.2015.00454>.
31. Tanaka D, Takahashi T, Yamashiro Y, Tanaka H, Kimochi Y, Nishio M, Sakatoku A, Nakamura S. Seasonal variations in bacterioplankton community structures in two small rivers in the himi region of central japan and their relationships with environmental factors. *World J Microbiol Biotechnol.* 2017;33(12):. <https://doi.org/10.1007/s11274-017-2377-4>.
32. Dudgeon D, Arthington AH, Gessner MO, Kawabata Z-I, Knowler DJ, Lévêque C, Naiman RJ, Prieur-Richard A-H, Soto D, Stiassny MLJ, Sullivan CA. Freshwater biodiversity: importance, threats, status and conservation challenges. *Biol Rev.* 2005;81(02):163. <https://doi.org/10.1017/s1464793105006950>.
33. Seto KC, Fragkias M, Gneralp B, Reilly MK. A meta-analysis of global urban land expansion. *PLoS ONE.* 2011;6(8):23777. <https://doi.org/10.1371/journal.pone.0023777>.
34. Waters CN, Zalasiewicz J, Summerhayes C, Barnosky AD, Poirier C, uszka AG, Cearreta A, Edgeworth M, Ellis EC, Ellis M, Jeandel C, Leinfelder R, McNeill JR, d Richter D, Steffen W, Syvitski J, Vidas D, Wagreich M, Williams M, Zhisheng A, Grinevald J, Odada E, Oreskes N, Wolfe AP. The anthropocene is functionally and stratigraphically distinct from the holocene. *Science.* 2016;351(6269):2622–2622. <https://doi.org/10.1126/science.aad2622>.
35. Isbell F, Gonzalez A, Loreau M, Cowles J, Diaz S, Hector A, Mace GM, Wardle DA, O'Connor MI, Duffy JE, Turnbull LA, Thompson PL, Larigauderie A. Linking the influence and dependence of people on biodiversity across scales. *Nature.* 2017;546(7656):65–72. <https://doi.org/10.1038/nature22899>.
36. Macher J-N, Leese F. Environmental DNA metabarcoding of rivers: Not all edna is everywhere and not all the time. 2017. bioRxiv. <https://doi.org/10.1101/164046>. <http://arxiv.org/abs/http://www.biorxiv.org/content/early/2017/07/15/164046.full.pdf>.
37. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Xu ZZ, Jiang L, Haroon MF, Kanbar J, Zhu Q, Song SJ, Kosciulek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Consortium TEMP. A communal catalogue reveals earth's multiscale microbial diversity. *Nature.* 2017. <https://doi.org/10.1038/nature24621>.
38. Power JF, Carere CR, Lee CK, Wakerley GL, Evans DW, Button M, White D, Climo MD, Hinze AM, Morgan XC, McDonald IR, Cary SC, Stott MB. Microbial biogeography of 1,000 geothermal springs in New Zealand. 2018. bioRxiv. <https://doi.org/10.1101/247759>. <http://arxiv.org/abs/https://www.biorxiv.org/content/early/2018/01/15/247759.full.pdf>.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)



## 2.2 PUBLICATION II

Theodor Sperlea, Nico Kreuder, Daniela Beisser, Georges Hattab, Jens Boenigk, and Dominik Heider (2021). Quantification of the covariation of lake microbiomes and environmental variables using a machine learning-based framework. *Molecular Ecology*, 30(9):2131–2144

**Contributions:** T. Sperlea designed and performed all data analyses, NK and DB contributed to the bioindicator analysis, DB and JB provided the data sets, JB, GH, and DH supervised the study. All authors discussed the results and wrote the manuscript.

\* \* \*

In microbial biomonitoring, the measured prevalences of certain microbial species act as a proxy measurement of other parameters of the ecosystem. These are, usually, ecosystem health parameters but can also include distinct physico-chemical parameters. A direct correlation between the latter and the prevalence can, however, be misleading. For one, the relation between the environmental parameter and the prevalence of the microbial species must be assumed to be indirect or, at least, greatly influenced by indirect effects. Furthermore, the response of one microbial species will be modulated by the prevalence of other, interacting microbial species.

In this publication, we used the status of the microbiome as a whole as a proxy measurement for a range of physico-chemical parameters. To that end, we developed a machine learning approach called the covariation framework. In it, a model is used to project the whole microbiome to a one-dimensional space comparable to the space the parameter in question is in. After that, we can calculate the  $R^2$  of the predicted and measured values of the parameter as a metric of covariation between the microbiome and the parameter in question. This methodology was applied to a large-scale microbiome dataset sampled from European lakes.

Evaluating different combinations of machine learning models, feature selection methods, and aggregation of the microbiome at different taxonomic levels, we found that Random Forest models combined with IndVal selection at the OTU level lead to the highest  $R^2$  values for parameters in general. This underscores (i) the non-linearity of interactions between microbial species, (ii) the biological relevance of the IndVal function, and (iii) the degree of loss of informativeness when aggregating microbes at lower levels of the taxonomy. Aside from reporting measures of covariation for a total of 27 parameters, we also present bioindicators for these and a list of multi-

task bioindicators, i.e., OTUs that appear as bioindicators for multiple environmental parameters.

\* \* \*

Beim mikrobiellen Biomonitoring dienen die erfassten OTU-Tabellen als Proxy-Messung für andere Parameter des Ökosystems. Dies sind in der Regel Parameter zur Gesundheit des Ökosystems, können aber auch bestimmte physikalisch-chemische Parameter umfassen. Eine direkte Korrelation zwischen letzteren und dem Vorkommen der Mikroorganismen kann jedoch irreführend sein. Zum einen muss davon ausgegangen werden, dass der Zusammenhang zwischen dem Umweltparameter und dem mikrobiellen Vorkommen indirekt ist oder zumindest stark durch indirekte Effekte beeinflusst wird. Zum anderen wird die Reaktion einer mikrobiellen Spezies durch die Prävalenz anderer, mit diesen interagierenden mikrobiellen Spezies moduliert.

In dieser Veröffentlichung haben wir den Status des Mikrobioms als Ganzes als Proxy-Messung für eine Reihe von physikalisch-chemischen Parametern verwendet. Zu diesem Zweck haben wir einen Ansatz des maschinellen Lernens entwickelt, der als *covariation framework* bezeichnet wird. Darin wird ein Modell verwendet, um das gesamte Mikrobiom auf einen eindimensionalen Raum zu projizieren, der mit dem Raum vergleichbar ist, in dem sich der betreffende Parameter befindet. Danach können wir den  $R^2$  der vorhergesagten und gemessenen Werte des Parameters als eine Metrik der Kovariation zwischen dem Mikrobiom und dem fraglichen Parameter berechnen. Diese Methodik wurde auf einen großen Mikrobiom-Datensatz angewendet, der aus europäischen Seen entnommen wurde.

Bei der Evaluierung verschiedener Kombinationen von maschinellen Lernmodellen, Methoden zur Merkmalsauswahl (*feature selection*) und der Aggregation des Mikrobioms auf verschiedenen taxonomischen Ebenen haben wir festgestellt, dass Random-Forest-Modelle in Kombination mit einer *feature selection* durch Bioindikatoridentifikation auf OTU-Ebene zu den höchsten  $R^2$ -Werten für Parameter im Allgemeinen führen. Dies unterstreicht (i) die Nicht-Linearität der Interaktionen zwischen mikrobiellen Spezies, (ii) die biologische Relevanz der IndVal-Funktion und (iii) den Grad des Verlustes an Informativität, wenn Mikroben auf niedrigeren Ebenen der Taxonomie aggregiert werden. Neben der Kovariation des Mikrobioms mit insgesamt 27 Parametern präsentieren wir auch Bioindikatoren für diese sowie eine Liste von Multi-Task-Bioindikatoren, d. h. OTUs, die als Bioindikatoren für mehrere Umweltparameter auftreten.

# Quantification of the covariation of lake microbiomes and environmental variables using a machine learning-based framework

Theodor Sperlea<sup>1</sup>  | Nico Kreuder<sup>2</sup> | Daniela Beisser<sup>2</sup> | Georges Hattab<sup>1</sup>  | Jens Boenigk<sup>2</sup> | Dominik Heider<sup>1</sup> 

<sup>1</sup>Faculty of Mathematics and Computer Science, University of Marburg, Marburg (Lahn), Germany

<sup>2</sup>Department of Biodiversity, Center for Water and Environmental Research, University of Duisburg-Essen, Essen, Germany

## Correspondence

Dominik Heider, Faculty of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, D-35032 Marburg (Lahn), Germany.  
Email: dominik.heider@uni-marburg.de

## Funding information

LOEWE, Grant/Award Number: MOSLA; Deutsche Forschungsgemeinschaft, Grant/Award Number: BO 3245/19-1.

## Abstract

It is known that microorganisms are essential for the functioning of ecosystems, but the extent to which microorganisms respond to different environmental variables in their natural habitats is not clear. In the current study, we present a methodological framework to quantify the covariation of the microbial community of a habitat and environmental variables of this habitat. It is built on theoretical considerations of systems ecology, makes use of state-of-the-art machine learning techniques and can be used to identify bioindicators. We apply the framework to a data set containing operational taxonomic units (OTUs) as well as more than twenty physicochemical and geographic variables measured in a large-scale survey of European lakes. While a large part of variation (up to 61%) in many environmental variables can be explained by microbial community composition, some variables do not show significant covariation with the microbial lake community. Moreover, we have identified OTUs that act as “multitask” bioindicators, i.e., that are indicative for multiple environmental variables, and thus could be candidates for lake water monitoring schemes. Our results represent, for the first time, a quantification of the covariation of the lake microbiome and a wide array of environmental variables for lake ecosystems. Building on the results and methodology presented here, it will be possible to identify microbial taxa and processes that are essential for functioning and stability of lake ecosystems.

## KEYWORDS

bioindicators, lake ecology, machine learning, microbial communities, microbial ecology

## 1 | INTRODUCTION

Anthropogenic changes to the environment are threatening the stability of ecosystems globally and contribute to unprecedented rates of species extinction with catastrophic consequences for life as we know it (Ceballos et al., 2015; Isbell et al., 2017; Steffen

et al., 2011; Tilman et al., 2017; Williams et al., 2015). To mitigate the destabilization and the collapse of ecosystems, we need a more refined understanding of how they function. Systems ecology offers a paradigm that describes ecosystems as dynamic and complex networks of interactions both among organisms as well as between the biotic and abiotic aspects of an ecosystem (Evans

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. *Molecular Ecology* published by John Wiley & Sons Ltd.



et al., 2013; Jørgensen, 2016; Otwell et al., 2018; Webster et al., 2018).

Through interactions and the flow of energy and nutrients, different parts of an ecosystem are connected. This is not limited to direct interactions as, for example, the number of predators in an ecosystem has both an effect on the number of prey as well as on the plants eaten by the prey (Krikorian, 1979; Ulanowicz, 2001). The dynamic adaptability of ecosystems to environmental changes makes it possible to identify bioindicators, i.e., organisms whose presence and prevalence can be used to estimate other variables of the ecosystem (Heink & Kowarik, 2010; Karimi et al., 2017).

Bioindicators are used in biosphere-based ecosystem monitoring schemes such as the ones implemented in European countries under the Water Framework Directive (Birk et al., 2012; Hering et al., 2010) but also hold insights into the autecology of organisms (i.e., their specific ecological needs and actions) as well as the functioning of an ecosystem as a whole (Plassart et al., 2019). This is the case since organisms will only emerge as indicative for environmental variables they respond to directly (because of their ecological niche) or indirectly (since they interact closely with organisms that are, in turn, responsive to changes in the respective environmental variable). Due to their functional diversity, high growth rates, large population sizes, and high surface-to-volume ratio, bacteria and microeukaryotes are very responsive to environmental changes and represent optimal bioindicators (Cordier et al., 2019; Frühe et al., 2020; Karimi et al., 2017; Merkley et al., 2004).

The advent of next-generation sequencing (NGS) has greatly facilitated the use of microbial bioindicators. Firstly, it made it possible to identify organisms based on their genetic makeup instead of visual features (Frühe et al., 2020; Kermarrec et al., 2014). Secondly, techniques such as amplicon sequencing have made it feasible to capture microbial community compositions present in environmental samples (Parks et al., 2017). As different microorganisms exhibit different responses to changes in an environmental variable, and these responses are modulated by other microorganisms, the microbial community composition as a whole is more indicative of the status of the ecosystem than a selection of bioindicator species separately.

However, while being rather intuitive, the systems ecology paradigm also exposes theoretical and methodical obstacles for the study of microbial communities. For example, the assumption of variable independence, which is a requirement for many statistical approaches, does not hold for all environmental variables or processes of an ecosystem. Similarly, in a system, processes are influencing and modulating each other, rendering the distinction between direct and indirect interactions hard or even infeasible (Jørgensen, 2016). This is especially the case for microbial communities, where interaction networks are hard to measure and validate (Cazelles et al., 2015; Harris, 2016; Heink & Kowarik, 2010; Röttgers & Faust, 2018a) and the distinction between indirect and direct interactions is an open question (Guimarães et al., 2017; Röttgers & Faust, 2018b). Indeed, many studies prove a high ecological relevance of indirect microbial interactions (Deltedesco et al., 2020; Miller & Travis, 1996).

Additional issues for the study of microbial communities stem from the sparsity and very high dimensionality of OTU tables (Röttgers & Faust, 2018b; Weiss et al., 2017). With a number of samples vastly lower than the number of regressors (in our case: taxa or OTUs), regression is ill-defined and the adjustment of the  $R^2$  value for the number of regressors is impossible. Usually, both the collection of more data as well as very stringent feature selection are suggested to counteract this. Both measures, however, are only of limited use for the study of microbial communities, as sampling and sequencing remain expensive and the high number of different microorganisms is a nonreducible property of the study object.

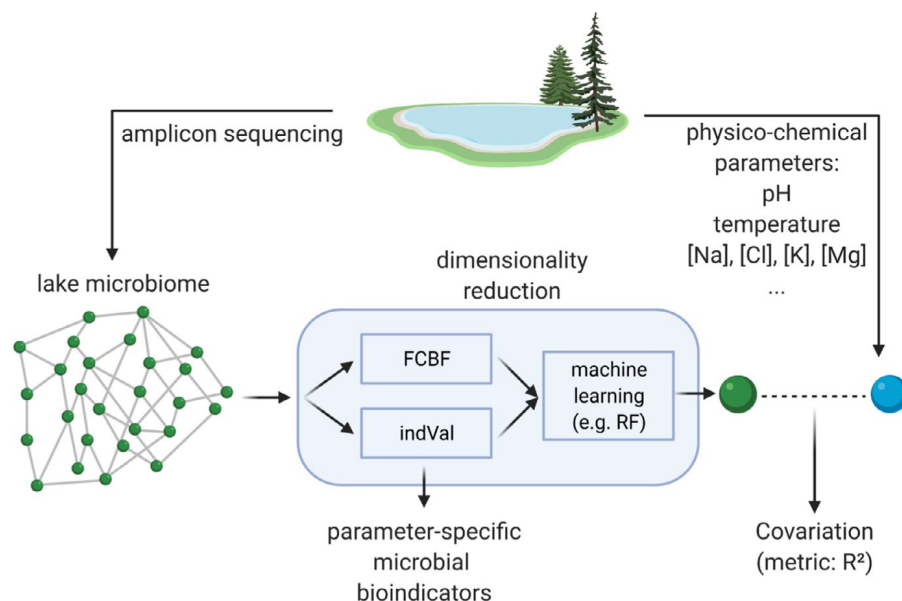
In this study, we developed methodological tools to study microbial communities in the context of systems ecology while acknowledging the aforementioned theoretical obstacles. Our main contribution is a machine learning-based framework for the quantification of the covariation between the microbiome and a total of 27 physicochemical and positional (i.e., GPS coordinates and altitude) variables of lake ecosystems (for an overview, see Figure 1, and for a list of variables, see Table 1). It builds upon a wealth of studies that elucidate the role of the microbiome in ecology using machine learning (Cordier, 2019; Cordier et al., 2018, 2019; Glasl et al., 2019; Grossmann, Beisser, et al., 2016; Han et al., 2019; Kiersztyn et al., 2019; Mikhailov et al., 2018; Sperlea et al., 2018; Tan et al., 2015). In our covariation framework, a machine learning model is trained to approximate a projection of the microbial prevalence space to a single dimension for each of the environmental variables, which makes it able to handle the extremely high dimensionality of amplicon-based microbiome data sets. The coefficient of correlation  $R^2$  between the projected microbial community composition and the measured environmental variable is, then, used as a metric of covariation. This corresponds to the covariation of the environmental variable and the whole microbiome, which is intuitively interpretable.

We applied this framework to a data set from a large-scale survey of European lakes (Bock et al., 2018, 2020; Boenigk et al., 2018). Lakes are considered as sentinels of ecosystem change at different temporal and geographical scales (García-García et al., 2019; Williamson et al., 2008). This is, in part, because lakes aggregate water from their catchments, and with it, pollutants and high nutrient concentrations. Furthermore, lakes are also directly affected by various anthropogenic stressors, such as overfishing, eutrophication, climate change, and invasive species (Dudgeon et al., 2005; World Wildlife Fund, 2018).

The use of nonlinear ensemble models facilitated a dimensional reduction of up to six orders of magnitude while retaining important relationships in the amplicon data set. Comparing two feature selection methods that are motivated by ecology, we found that filtering for bioindicators leads to a favourable behaviour of the framework. Analysing the operational taxonomic units (OTUs) identified as bioindicators in the feature selection step, we identified bacteria and microbial eukaryotes indicative of multiple environmental variables of lakes, which support the notion of high interdependency between ecological variables.

At the time of writing and to our knowledge, we provide the first large-scale, sequencing-based analysis of the potential of the full





**FIGURE 1** Graphical summary of the machine learning approach presented in this paper. Using amplicon sequencing data as well as a set of environmental variables collected in a large-scale survey of European lakes (Bock et al., 2018, 2020; Boenigk et al., 2018), we developed a novel approach to measuring the covariation of the whole microbiome and the environmental variables of lake ecosystems. At its core, our framework makes use of supervised machine learning methods to reduce the dimensionality of the microbial community composition. As the environmental variables are numerical features, the  $R^2$  metric serves as metric for covariation. Two different feature selection methods were compared, and microbial bioindicators were extracted from the IndVal method. Created with BioRender.com

**TABLE 1** Number of operational taxonomic units (OTUs) identified as bioindicators in the IndVal analysis for environmental variables of lakes. This list does not include the positional variables Coord.O and Coord.N as the interpretation of bioindicators for these is not clear

Variable	Number	Variable	Number
Alk. Gran	16	HCO <sub>3</sub>	27
Altitude	1595	K <sup>+</sup>	118
Anions	89	LF	1349
Ca <sup>2+</sup>	32	Mg <sup>2+</sup>	70
Cations	164	Na <sup>2+</sup>	86
CatSum	5	NH <sub>4</sub>	6
Cl <sup>-</sup>	48	NO <sub>3</sub>	17
COND	1	pH	603
DN	0	SO <sub>4</sub>	3
DP	0	Sumlons	166
DOC	43	T	920
DRSi	1	TP	92
H <sup>+</sup>	15		

microbial community composition as an indicator for physicochemical variables in lake ecosystems. To that end, we report a comprehensive quantification of the covariation of the complete microbiome with regard to these environmental variables. Our results highlight the advantages of machine learning methods for the study of microbial communities in a systems ecology paradigm. Furthermore, they underscore the importance of including bacteria and microeukaryotes at the species or OTU level into ecological monitoring schemes.

This work paves the way for future endeavours to better uncover the functional workings of ecosystems.

## 2 | MATERIALS AND METHODS

### 2.1 | Sample collection

Sampling was part of a pan-European study conducted in August 2012 (eukaryotic sequences are published in Boenigk et al., 2018; NCBI Bioproject PRJNA414052, prokaryotic sequences are published and described in Nuy et al., (2020), Bock et al., (2020); NCBI Bioproject PRJNA559862). To analyse the effects of biogeochemical factors on bacterial and protist freshwater communities on a large scale, 280 lakes were sampled, covering a broad latitudinal gradient ranging from Spain to the South of Scandinavia and altitudes from sea level to 3110 m.a.s.l. The samples were taken in daylight from the shore of each lake or pond collecting epilimnial water up to 0.5 m depth. Sampling details and information on measured physicochemical and geographic factors can be found in Boenigk et al., 2018. For DNA analyses filtered samples were air-dried and frozen in liquid nitrogen (Cryoshippers) and stored at  $-80^{\circ}\text{C}$  until further processing.

### 2.2 | DNA extraction and sequencing

Genomic DNA was extracted using the my-Budget DNA Mini Kit (Bio-Budget Technologies GmbH, Krefeld, Germany) following the protocol of the manufacturer and modifications after Boenigk et al.,

2018. Bacterial amplicon sequencing targeted the V2-V3 region of the 16S rRNA gene, eukaryotic amplicon sequencing targeted the V9 region of the 18S, and the ITS1 gene in the SSU genomic region. Samples were commercially sequenced (Fasteris, Geneva, Switzerland) using paired-end Illumina HiSeq 2500 sequencing in the “rapid run” mode to generate  $2 \times 300$  bp reads. For details, please see Boenigk et al., (2018), Nuy et al., (2020) and Bock et al., (2020).

## 2.3 | Sequence analysis

Adapter removal, quality trimming, and demultiplexing using index sequences were performed by the sequencing company (Fasteris). Sequence processing was performed using a provisional version of the Natrix pipeline (Welzel et al., 2020). Base quality of raw sequence reads was rechecked using the FASTQC software (v0.11.5; Andrews, 2010) and reads with an average Phred quality score below 25 or with at least one base with a Phred quality score below 15 were removed using PRINSEQ-LITE (v0.20.4; Schmieder & Edwards, 2011). The paired-end reads were assembled and quality filtered with the tool PANDASEQ (v2.10; Masella et al., 2012). Reads with uncalled bases, an assembly quality score below 0.9, a read overlap below 20 bases, or a base with a recalculated Phread-score below 1 were discarded. Assembled sequences were dereplicated and chimeras identified using UCHIME (usearch v7.0.1090; Edgar et al., 2011). Additionally, a split-sample filtering protocol (AmpliconDuo; Lange et al., 2015) was used to discard sequences that were not found in both technical replicates (A and B variant). Remaining sequences were clustered using SWARM (v2.2.2; Mahé et al., 2014) and OTU tables were generated based on this clustering. The eukaryotic representative sequences were further clustered by identical V9 sequences (V9\_Clust.R; Jensen, 2017). The taxonomic assignment of the eukaryotic sequences was performed by searching the NCBI database using BLAST (BLAST +v2.7.1; NCBI nt sequences from Dec 5, 2017). For the prokaryotic sequences SILVA (SILVA SSURef release 132) was used.

## 2.4 | Data preparation

Values for temperature (T) and conductivity (LF), measured in field in triplicates, were averaged for each sample. For the analyses at different taxonomic levels, for each taxon at each taxonomic level, OTU counts belonging to this taxon were aggregated. OTUs missing a taxonomic annotation at a taxonomic level were not counted.

To circumvent the problem of missing values in the environmental parameter data set, two subdata sets were created, namely the all\_samples and all\_features subdata sets. The all\_samples subdata set contains the environmental variables measured in the field (altitude, GPS coordinates, pH, conductivity, temperature, and time of sampling) and OTUs for 241 lakes. An additional set of 21 physicochemical variables had been measured for a subset of 47 lakes. Excluding the positional variables and the time measurement, lakes

with the extended feature set and the corresponding OTUs constitute the all\_features data set.

Outliers in the environmental variables were defined as data points falling outside of a range of 1.5 times the interquartile range below the first or above the third quartile (as calculated using the R function `boxplot.stats()`). Samples that contain at least one outlier in any of the environmental variables relevant for the subdata set were excluded from further analysis, leading to 201 and 42 samples in the all\_samples and all\_features data set, respectively (see Table S1 for a list of lakes present in the subdata sets). This was done to reduce the variability in the data set as well as to remove potential measurement errors. OTUs and taxa absent from all samples in one of these subdata sets were removed. OTU and taxon counts were centred and scaled using the R function `scale()` before training.

## 2.5 | Covariation framework and machine learning

At the core of the covariation framework is a model that is trained to approximate this environmental variable based on the OTU table or taxonomically aggregated prevalence table. In the covariation framework, however, the common supervised machine learning approach is interpreted in a novel way, that is consistent with the theory behind machine learning as well as systems ecology: The prediction of the framework is interpreted as a projection of the microbial community composition to a single dimension that is comparable to the environmental variable the model was trained on. As metric for covariation, the coefficient of determination  $R^2$  between the dimensionality-reduced microbiome and the measured values of the variable for the held-out samples was used. As a secondary metric, the root-mean-square error was also calculated.

The full model used to quantify the amount of covariation of the microbial community and an environmental variable consists of a feature selection method and a machine learning model, both of which will be described in the following paragraphs in more detail. These two steps form the full model of the covariation framework and are evaluated as one, for example the feature selection as well as the machine learning are evaluated based on the full model performance. Because of the low number of samples analysed here, a cross-validation scheme was used for model training and prediction as this results in final models with low bias even for small data sets (Bishop, 2006). This will be described at the end of the subsection.

Feature selection was performed using either a fast correlation-based filter (FCBF) (Yu & Liu, 2003) or the `multipatt()` function (IndVal method, 999 random permutations) from the R package `indicspecies` (v1.7.9; Cáceres & Legendre, 2009). The choice of the former was motivated by the widespread use of correlation networks as proxies for microbial interactions (Proulx et al., 2005). In these, nodes represent species and are connected with an edge if their prevalence correlates across a range of samples. Along these lines, FCBF

groups OTUs or taxa that are neighbours in a correlation network into syntaxa, i.e., groups of organisms that act as one unit in environmental changes (Chaffron et al., 2010). For this filter, a cutoff of 0.6 was chosen for the Pearson correlation coefficient, because, consistently over taxonomic levels, only around one percent of intertaxa correlations showed higher correlation coefficients. For the IndVal analysis, which is of widespread use in ecological studies, samples were separated by tertiles of the variable in question and OTU and taxon occurrence numbers were standardised using the Hellinger transformation to decrease the influence of highly abundant OTUs (Legendre & Gallagher, 2001).

A total of seven machine learning models from the R package caret (v6.0.86; Kuhn, 2008) were used as base learners in this study: random forest (rf), stochastic gradient boosting (gbm), extreme gradient boosting (xgbTree), support vector machines with linear and radial kernel (svmLinear, svmRadial), generalised linear model (glmnet), and k-nearest neighbours (knn). These models were trained using the train() function with default parameters, which includes hyperparameter tuning by grid search. Model predictions were generated using the predict() function.

To use a cross-validation scheme, the full set of samples was split into  $k$  subsets of approximately equal size, and  $k$  models are trained and used for prediction separately. While higher values of  $k$  are known to reduce the bias in the evaluation, the runtime of the whole training process is also greatly affected by  $k$ . Thus, for the all\_samples subdata set a 10-fold cross-validation and for the all\_features subdata set a leave-one-out cross-validation scheme was used as follows. For fold  $i$ , all subsets except for subset  $i$  were used in the training phase. The training phase consisted in, firstly, feature selection of the input features (i.e., taxa and OTU tables), and, secondly, fitting a model to approximate the target variable based on the selected features. Then, the fitted model was used to predict the target variables based in the held-out subset  $i$ . These predictions were collected and compared to the respective measured value of the environmental variable in question. As performance metrics, the coefficient of determination  $R^2$  and root-mean-square error were calculated using the postResample function from the R package caret (v6.0.86; Kuhn, 2008). Confidence intervals for performance metrics were determined by 1000× resampling by bootstrapping of predicted and measured values.

## 2.6 | Data analysis

Environmental variables were clustered according to their Pearson correlation using the hclust() function from the R package stats (v4.0.1). Variable importances were extracted from rf models using the varImp() function from the R package caret (v6.0.86, Kuhn (2008)) and averaged over the training folds. The ttest() function was used to assess whether there is a difference in the feature importance of eukaryotic and prokaryotic OTUs, and the resulting  $p$ -values were adjusted using the Benjamini-Hochburg method as implemented in the p.adjust() function.

## 3 | RESULTS

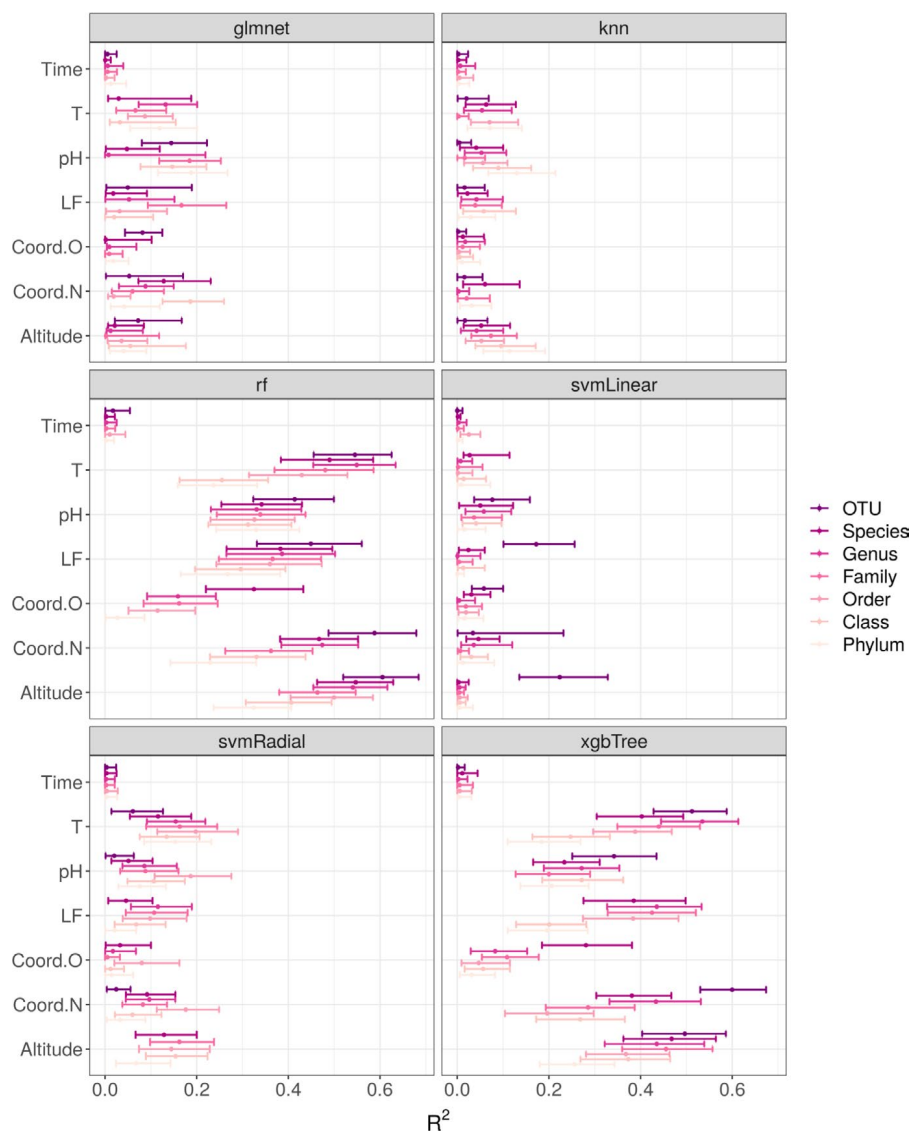
### 3.1 | Nonlinear models capture relevant patterns in microbial community composition

In general, regression models can be seen as approximating a function that projects the input feature space to a one-dimensional target space, thus performing a supervised dimensionality reduction. Based on this notion, we developed a framework to quantify the covariation of an ecosystem's microbial community composition and an environmental variable. In it, we train supervised machine learning models to project the OTU- or taxonomy-table microbiome obtained by amplicon sequencing to a single dimension that is comparable to the target variable in question. As a metric of covariation, we used the coefficient of determination,  $R^2$ , calculated between the one-dimensional microbiome (i.e., the prediction in a traditional machine learning scheme) and the measured values of the environmental variable values. We also calculated the root-mean-square error as a secondary metric for the performance of the framework (see Table S2). However, in the context of the covariation framework, the  $R^2$  metric lends itself to a more straightforward interpretation, i.e., as the amount of variation in the environmental variable explained by the projected, one-dimensional microbial community composition.

In a first implementation of the framework, we employed a fast correlation-based filter (FCBF) to reduce the dimensionality before machine learning, and trained machine learning models on the all\_samples data set (and, therefore, only for a reduced number of variables) using a 10-fold cross-validation evaluation scheme. The choice of this feature selection method was motivated by the use of correlation networks for microbial communities (Proulx et al., 2005).

To test the hypothesis that nonlinear, as well as linear, relationships between microorganisms are important for their response to environmental changes, we compared the performance of different regression models. Higher  $R^2$  values indicate a higher propensity of the model to capture relevant patterns in the microbial community composition. In our results, models that can approximate both linear and nonlinear relationships between features (i.e., Random Forest and xgbTree) outperform linear models. This result suggests that nonlinear projections are necessary to capture environmentally relevant patterns in the microbiome in a single dimension and thus supports the notion that complex relationships are present between microbial community structure and environmental variables (see Figure 2). Based on this, we focus the presentation and discussion of further results to Random Forest models.

However, FCBF does not reduce the dimensionality of the microbial community composition sufficiently to enable the training of regression models for all levels. Especially at the OTU level, around 89% of the initial features were still left after feature selection (Table 2). This disproportion between sample number and feature space (i.e., taxon or OTU table) dimensionality made the application of the framework impossible for some of the environmental variables (see missing values in Figure 2).



**FIGURE 2** Covariation of the microbial community composition of a lake and its variables, for the all\_samples data set using fast correlation-based filter (FCBF) as feature selection method. Lines represent 95% confidence intervals calculated from resampling, dots represent the median of resampled values. Some of the model-variable combinations are not computable because of too high microbial community dimensionality

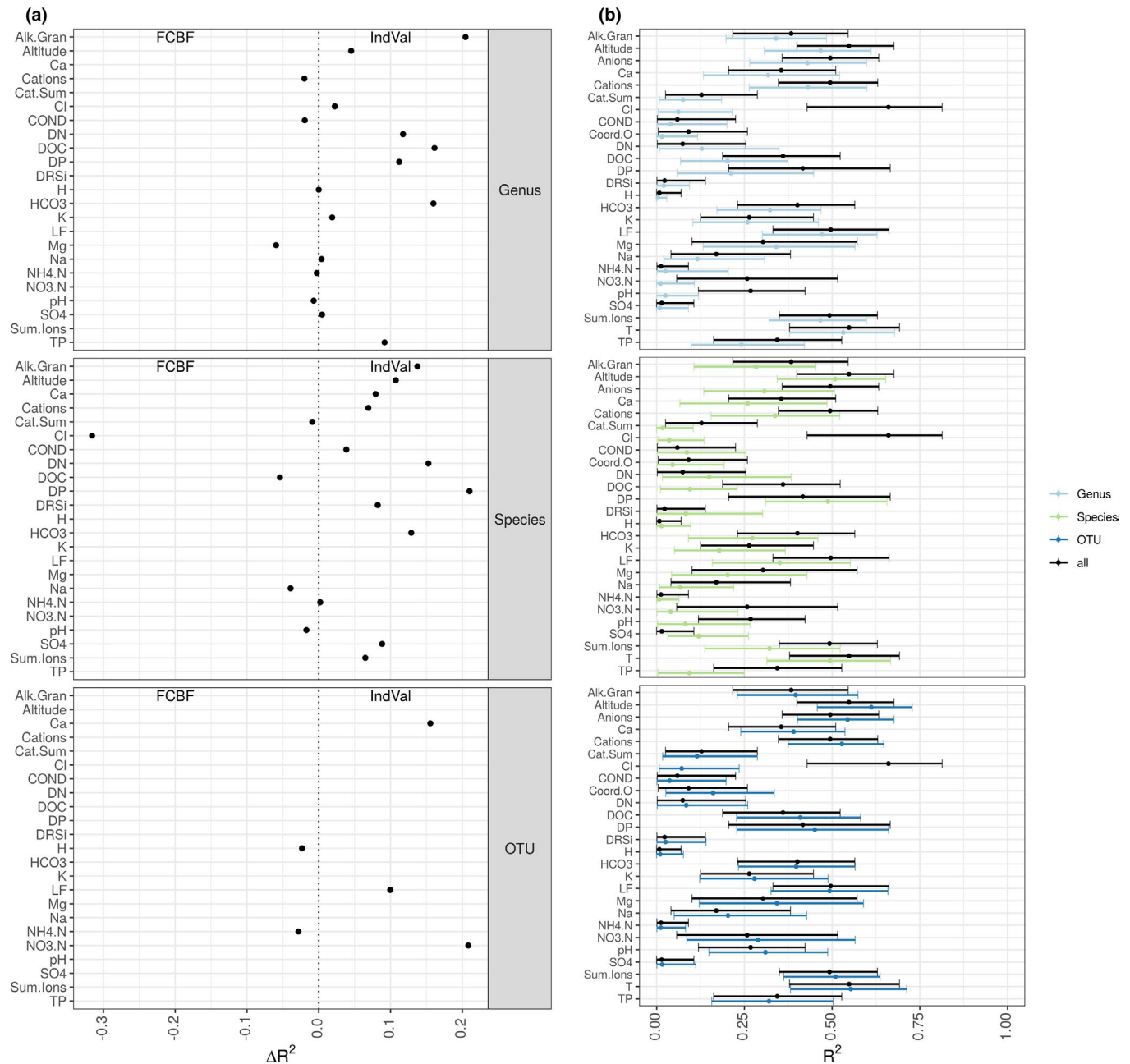
**TABLE 2** Dimensionality of taxonomic levels, as well as average dimensionality after dimensionality reduction via fast correlation-based filter (FCBF) and the IndVal method for the all\_features data set

Level	Taxa	FCBF	IndVal
Domain	3	3	–
Phylum	76	76	3.22
Class	253	244	10.10
Order	752	714	24.16
Family	885	857	33.72
Genus	2 353	2242	69.41
Species	5 384	4967	80.49
OTU	315,731	279,952	721.07

### 3.2 | Indicator species analysis as feature selection for microbiome dimensionality reduction

As an alternative filtering method, we employed the IndVal method (Dufrêne & Legendre, 1997). This calculates a composite indicator

value based on the specificity and fidelity of a given species concerning a predefined set of sites. Its use in the identification of bioindicators suggests that it should be able to select OTUs or taxa that covary with a given lake variable. Applying IndVal as a feature selection method in our framework to the all\_features data set resulted in more stringent models for OTUs and taxa (Table 1). Comparing the results of the framework developed earlier using either FCBF or IndVal as feature selection method shows that, for some environmental variables and taxonomic levels, using IndVal leads to better results, albeit not significantly (see Figure 3a). Furthermore, for some combinations of taxonomic levels and environmental variables, the use of FCBF outperformed the use of IndVal. On the other hand, while some FCBF runs were not computable (highlighted by the missing values in Figure 3a), this was never the case for IndVal runs. Finally, as the models trained using IndVal selected features are more sparse, this filter method is, in general, preferable to FCBF for microbial communities. Based on these results, we conclude that most of the taxa or OTUs that covary with the respective environmental variable are contained in the IndVal selection.



**FIGURE 3** IndVal as feature selection method for the all\_features data set. (a) Difference of median  $R^2$  between models trained on fast correlation-based filter (FCBF)- and IndVal-filtered microbial community composition. Negative values indicate better performance using FCBF, positive values indicate better performance using IndVal. Missing data points indicate combinations of taxonomic level and environmental variable that were not computable because of too high dimensionality after using FCBF. (b) Quantification of covariation of the microbial community composition and physicochemical variables of a lake using IndVal as feature selection method. Lines represent 95% confidence intervals calculated from resampling, dots represent the median of resampled values. Grey lines (labelled "all") represent results of models trained on a concatenation of all data from different taxonomic levels. Ion names represent concentrations. For the results for other models and taxonomic levels, see Table S2

### 3.3 | Covariation at different taxonomic levels

Random Forest models trained with IndVal-selected features at the OTU level lead to median  $R^2$  values of more than 0.3 for more than half of the physicochemical variables present in the all\_features data set (Figure 3b). As seen for FCBF (see Figure 2), lower taxonomic levels covary more with the physicochemical variables than do higher

levels. However, the results from different levels of taxonomy should be compared with care, because the number of regressors (i.e., taxa or OTUs) increases strongly with falling taxonomic level and the  $R^2$  is known to increase monotonically with the number of regressors. The usual way to alleviate this is to adjust  $R^2$  values to the feature space dimensionality, but this is not possible here because the dimensionality is much higher than the numbers of samples of the analysed data sets.



To test the hypothesis that different levels of microbial taxonomy respond with environmental variables in different ways, we aggregated the IndVals over different levels of taxonomy and used this data to train machine learning models (lines labelled “all” in Figure 3b). These models do not significantly outperform the models trained on OTU prevalence tables although they were trained with a higher number of regressors (i.e., the sum of all taxa and OTUs). Therefore, we conclude that higher taxonomic levels do not contribute to ecologically relevant patterns not already present at the OTU level.

### 3.4 | Analysis of microbial multitask bioindicators

The results presented to this point support the use of the IndVal to identify ecologically relevant taxa and OTUs from amplicon sequencing data. The numbers of bioindicators for different variables at different levels of taxonomy obtained this way ranged over four orders of magnitude (see Table 1 and Table S3). We analysed these bioindicator OTUs by focusing on multitask bioindicators, i.e., OTUs that emerged as indicative for multiple environmental variables and might, therefore, act as general indicators of lake ecosystem status.

All of the bioindicators indicative of more than seven variables are annotated as Bacteria (see Table 3 and Figure 4a) except for two OTUs that are annotated as chloroplasts of the green algae *Phacotus lenticularis*. This organism has been described as a bioindicator for freshwater ecosystems before (Jiang & Shen, 2005; Schlegel et al., 1998). Most of the other OTUs are from the Phyla Bacteroidetes and Proteobacteria. Many of the lowest distinct taxa we identified have previously been discussed as bioindicators for general ecosystem quality (Ignavibacteriales (Cordier, 2019), *Limnobacter* (Yang et al., 2019), and Sandaracinaceae (Wei et al., 2019)), certain environmental variables (*Opitutus* (Puranik et al., 2016; Plassart et al., 2019), Alcaligenaceae (Sharuddin et al., 2017), *Novosphingobium* (Astudillo-García et al., 2019; Reis et al., 2020), and NS11-12 marine group (Coclet et al., 2019; Henson et al., 2018)), and human interference/impact/pollution (*Actibacter* (Kegler et al., 2018), *Fluviicola* (Chen et al., 2019), and SC-I-84 (Pershina et al., 2015)). However, not all of these taxa have previously been identified in lake ecosystems, and most of the OTUs among these bacterial multitask bioindicators are assigned to taxa originally isolated from soil ecosystems (see Table 3).

The multitask bioindicators among the eukaryotes are, at most, indicative for five environmental variables. Among the 32 OTUs that are indicative for more than two variables, six are annotated as Ciliophora or Chlorophyta. These classes are ubiquitous in lakes (Grossmann, Jensen, Heider, et al., 2016; Grossmann, Jensen, Pandey, et al., 2016; Mikhailov et al., 2018), contain many species that inhabit specific ecological niches and have been used as bioindicators (Bellinger & Sigee, 2015; Foissner & Berger, 1996; Lee et al., 2004). Similarly, many of the eukaryotic multitask OTUs identified here belong to genera that have been described as ubiquitous in freshwater ecosystems (e.g., Chytridiomycota (Bai et al., 2018),

*Desmodesmus* (Johnson et al., 2007) or *Gymnodinium* (Thessen et al., 2012)). However, most of the species we identified have, to our knowledge, not yet been described as bioindicators at lower taxonomic levels. Notably, clustering the environmental variables according to their pairwise Pearson correlation results in patterns of multitask bioindicators in Figure 4. This further supports the notion that an interaction network underlies the microbial community structure of lake ecosystems and this network is shaped by environmental variables.

Based on our finding that bacterial OTUs can be indicative for more than five environmental variables at the same time, we speculated that bacteria are, in general, better suited as bioindicators than eukaryotes. To test this hypothesis, we extracted feature importance values from the Random Forest models used in the covariation framework. With the null hypothesis that the feature importances of eukaryotes and prokaryotes have the same means, we ran two-sample *t* tests and found significant differences ( $p < .05$  after Bonferroni-Hochberg correction with  $n = 23$ ) for the bioindicators for altitude, dissolved organic carbon (DOC), dissolved reactive silica (DRSi), hydrogen (H), potassium (K), ammonium ( $\text{NH}_4$ ), nitrate ( $\text{NO}_3$ ), sum of ions (Sum. Ions), and temperature (T) (see Table S5). For these environmental variables, thus, the mean feature importance of eukaryotic and prokaryotic bioindicators can be regarded as different. This result suggests that, at least with regard to these variables, bacteria and eukaryotes play different roles in lake ecosystems. However, for the other variables, we observed no significant difference in feature importances between bacterial and eukaryotic OTUs. This supports the notion that in an ecosystem, groups of interacting organisms cannot be seen as fully independent with regards to their ecological function.

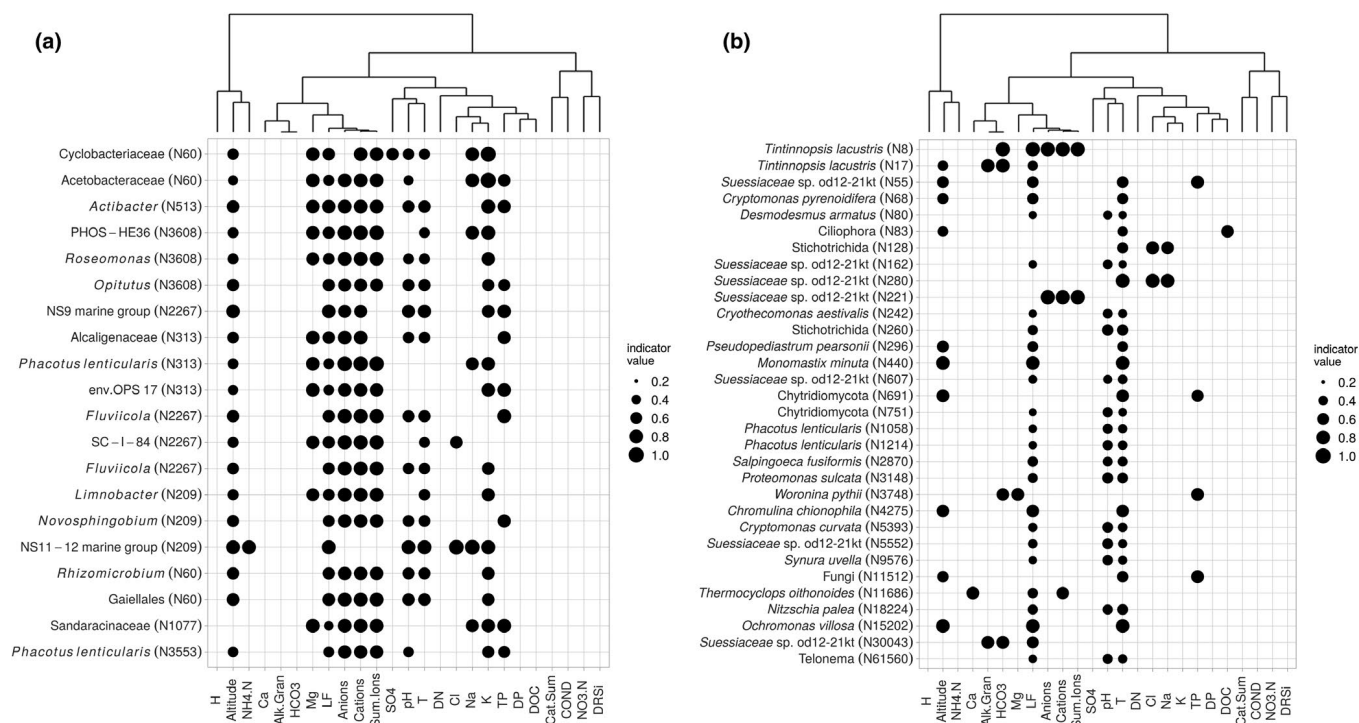
## 4 | DISCUSSION

To arrive at a fuller image of the functioning of ecosystems, methodological approaches and theoretical paradigms have to be integrated. In this study, we combined bioindicator analysis, machine learning techniques, and the systems ecology paradigm to quantify the covariation of the microbiome and environmental variables of lake ecosystems. We present a framework that acknowledges the technical obstacles presented by ecological data in general and molecular microbial community data sets in particular.

For the design of the covariation framework, we compared different machine learning models and found that ensembles of decision trees (such as Random Forest and xgbTree models) were best able to project the microbiome to a one-dimensional space as judged by the  $R^2$  metric (Figure 1). This is most probably due to their ability to approximate highly nonlinear relationships and cope with large feature spaces (Breiman, 2001). Additionally, ensembles of decision trees are, in principle, capable of learning from data for which the independence assumption does not hold (Breiman, 2001). We were also able to show that while using FCBF and IndVal as feature selection methods leads to comparable results, the IndVal method results

TABLE 3 Multitask bioindicators that have been identified for more than seven environmental variables. Highlighted rows contain chloroplasts identified based on 16 s rRNA sequence. For a overview of variables and indicator statistic for each of these operational taxonomic units (OTUs), see Figure 4

ID	Freq	Phylum	Class	Order	Family	Genus	Species
N1077	10	Bacteroidetes	Cytophagia	Cytophagales	Cyclobacteriaceae	Uncultured	Uncultured bacterium
N3553	10	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Roseomonas	Roseomonas sp. S08
N513	10	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Actibacter	Uncultured bacterium
N2267	9	Ignavibacteriae	Ignavibacteria	Ignavibacteriales	PHOS-HE36	Uncultured soil bacterium	
N2497	9	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Roseomonas	Groundwater biofilm bacterium H2
N569	9	Verrucomicrobia	Opitutae	Opitales	Opitutaceae	Opitutus	Uncultured soil bacterium
N177	8	Bacteroidetes	Flavobacteriia	Flavobacteriales	NS9 marine group	Uncultured bacterium	
N1886	8	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Uncultured	Uncultured soil bacterium
N209	8	Chloroplast of Phacotus lenticularis					
N2139	8	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	env.OPS 17	Uncultured bacterium	
N313	8	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cryomorphaceae	Fluviicola	Uncultured bacterium
N3608	8	Proteobacteria	Betaproteobacteria	SC-I-84	uncultured bacterium		
N395	8	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cryomorphaceae	Fluviicola	Uncultured Bacteroidetes bacterium
N426	8	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Limnobacter	Uncultured bacterium
N533	8	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	Uncultured bacterium
N60	8	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	NS11-12 marine group	Uncultured Sphingobacterium sp.	
N636	8	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiales Incertae Sedis	Rhizomicrobium	Uncultured bacterium
N642	8	Actinobacteria	Thermoleophilia	Gaiellales	uncultured	uncultured bacterium	
N6836	8	Proteobacteria	Deltaproteobacteria	Myxococcales	Sandaracinaceae	uncultured	Uncultured bacterium
N735	8	Chloroplast of Phacotus lenticularis					



**FIGURE 4** Multitask bioindicators for lake ecosystem variables. (a) Operational taxonomic units (OTUs) indicative for more than seven variables, (b) Eukaryotic OTUs indicative for more than two variables. Dot size represents indicator statistic magnitude. Dendrogram and variable order are derived from all-vs.-all Pearson's correlation in the all\_features data set. For taxonomic annotation of the OTUs, see Table 3 and Table S4, for (a) and (b), respectively

in sparser models that allow the use of the framework even for extremely high-dimensional data sets at low levels of the taxonomy (see Table 2). While IndVal has been used for molecular data sets collected, for example, at the Great Barrier Reef (Glasl et al., 2019), this study is first in applying it to molecular data in the context of lake ecology.

Applying our framework to a data set of microbial communities collected in a large-scale survey of European lakes, we were able to quantify the covariation between the lake microbiome and a list of environmental parameters for different levels of taxonomy (Figure 3b). Due to the high dimensionality of environmental microbiomes, we are not able to conclude whether OTUs show a covariation significantly higher than any of the other levels of taxonomy. Nevertheless, our results show that, for most environmental variables, higher levels of taxonomy do not contain relevant patterns not already present on the OTU level (see Figure 3b), which is in contrast to the findings of others (Washburne et al., 2017).

In the analysis of bioindicator OTUs identified in this study, we focused on multitask bioindicators. Among the OTUs identified as bioindicators for more than seven environmental variables, most have been taxonomically assigned to uncultured soil bacteria (see Figure 4a and Table 3). Similarly, most high-ranked eukaryotic multitask bioindicators (see Figure 4b and Table S4) have been first identified in freshwater biomes, but not necessarily been found in lake samples yet. As the data set analysed here stems from lakes, this is most probably an artefact of imprecise taxonomic annotation (Chen et al., 2013), but might also point to the diversity of ecological niches

inhabited by bacterial subspecies grouped into one OTU or species (García-García et al., 2019). Although soil and freshwater microbial community compositions differ significantly (Grossmann, Jensen, Heider, et al., 2016), microorganisms can enter lakes from soil ecosystems directly or, for example, via rivers that feed the lake. The emergence of *Phacotus lenticularis* as a multitask organism in both groups of organisms (see Table 3 and Figure 4b) underscores its role as a bioindicator.

Recent studies have argued for differences in ecological function between bacteria and microbial eukaryotes in lake ecosystems (Bock et al., 2020; Logares et al., 2018; Massana & Logares, 2012). More specifically, it has been argued that bacteria are more responsive to environmental changes than eukaryotes (Bock et al., 2020; Frühe et al., 2020; Karimi et al., 2017; Merkley et al., 2004). This is supported by the results of our study that bacteria that are multitask bioindicators can be indicative of more environmental variables of lakes than eukaryotic multitask bioindicators (see Figure 4). Moreover, we also found significant differences in the variable importances assigned to bacterial and eukaryotic OTUs by the Random Forest model used in the framework for some environmental variables (see Table S5). However, this is not the case for all variables. There are two main reasons for this. First, at the domain level, aggregated prevalence numbers do not covary much with environmental variables (see Figures 2, 3b). Second, the interactions between organisms lead to indirect effects that would inhibit such a simple distinction between eukaryotes and bacteria. In the context of systems ecology, we would not expect



groups of organisms to be independent in a manner relevant to this question.

Unsurprisingly, the variables these multitask bioindicators are indicative of show a high degree of correlation (see Figure 4). Aside from underscoring the need for functional diversity in bioindicators if aiming at covering all environmental variables, this indicates that there are “main factors” among lake variables that influence a high number of other variables strongly. Altitude has been described as one of them, as it is directly or indirectly related to, among others, temperature, radiation, salinity, conductivity, and nutrient concentration (Karlsson et al., 2005). This is the case because lakes in the lowland mainly arise from rivers that have their source in mountain chains and get enriched with nutrients during their courses. In particular for eukaryotic multitask bioindicators, our analyses suggest that temperature, conductivity (as measured in the field, displayed in this study under the label “LF”), and pH might also act as “main factors”.

Nevertheless, our results also underscore the need for further studies that include large-scale amplicon sequencing surveys of ecosystems. This is mainly the case because the natural variability of environmental samples in general and lake ecosystems specifically is very high, leading to the rather large confidence intervals observed in this study. Thus, including more samples in analyses such as ours would enable to better model the heterogeneity of natural ecosystems and lead to more robust and powerful statistical results. Further studies that are based on larger data sets should also allow for analyses based on less-stringent outlier removal than applied in this study, representing a wider array of natural variation of lake ecosystems.

In principle, both the covariation framework as well as the bioindicator analysis can easily be applied to samples from different environmental sources and other sequencing methods, such as metagenomic and metatranscriptomic data sets, as long as there is a straightforward interpretation for the results of the IndVal method. This is especially noteworthy as the importance and popularity of metagenomic assays in microbial ecology has risen fast in the last years (Awasthi et al., 2020; Hugerth et al., 2015; Panwar et al., 2020; Vishnivetskaya et al., 2020; Zeng et al., 2016).

Taken together, our results represent an important contribution to the discussion around the use of microorganisms in lake ecosystem monitoring schemes. First, they indicate that the physicochemical status of a lake cannot fully be predicted by its microbiome (see Figures 2, 3b). Nevertheless, up to around 60% of the variation in certain variables can be predicted by the lakes microbial community composition, which is comparable to results from soil ecosystems (Hermans et al., 2016). Second, the predominance of bacteria among multitask bioindicators (see Figure 4) supports the view that, in lake ecosystems, bacteria are more responsive to changes in environmental variables than eukaryotes (Bock et al., 2020). This underscores the importance of including prokaryotes into official ecosystem monitoring schemes. Third, our results offer an insight into the autecology of microbial taxa and OTUs in their natural habitats by indicating which microbes react strongly to changes different environmental variables. These insights can lay the groundwork for

novel, niche-based analyses of environmental microbiomes (Chase & Leibold, 2003).

## ACKNOWLEDGEMENTS

Calculations on the MaRC2 high-performance computer of the University of Marburg were conducted for this research. We would like to thank René Sitt of HPC-Hessen, funded by the State Ministry of Higher Education, Research and the Arts, for installation and maintenance of software on the MaRC2 high-performance computer. We would like to thank Julia Nuy and Marius Welzel for helping with data availability. This work was supported by the LOEWE program of the State of Hesse (Germany) in the MOSLA research cluster. We also acknowledge funding by the Bauer-Foundation and Stemmler-Foundation for the project “Differential potential of metabarcoding, metatranscriptomics, and metagenomics for the assessment of lake water quality” and of the DFG project BO 3245/19-1. Open Access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST

The authors declare that they have no competing financial interests.

## AUTHOR CONTRIBUTIONS

Theodor Sperlea designed and performed the data analyses, Nico Kreuder and Daniela Beisser contributed substantially to the bioindicator analysis, Daniela Beisser and Jens Boenigk provided the data sets, Jens Boenigk, Georges Hattab, and Dominik Heider supervised the study. All authors discussed the results and wrote and revised the manuscript.

## DATA AVAILABILITY STATEMENT

Raw sequencing data are available under the NCBI BioProject IDs PRJNA414052 and PRJNA559862, and the physico-chemical parameter data under the <https://doi.org/10.6084/m9.figshare.14039312>.

## ORCID

Theodor Sperlea  <https://orcid.org/0000-0003-4307-2963>

Georges Hattab  <https://orcid.org/0000-0003-4168-8254>

Dominik Heider  <https://orcid.org/0000-0002-3108-8311>

## REFERENCES

- Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Astudillo-García, C., Hermans, S. M., Stevenson, B., Buckley, H. L., & Lear, G. (2019). Microbial assemblages and bioindicators as proxies for ecosystem health status: potential and limitations. *Applied Microbiology and Biotechnology*, 103(16), 6407–6421.
- Awasthi, M. K., Ravindran, B., Sarsaiya, S., Chen, H., Wainaina, S., Singh, E., Liu, T., Kumar, S., Pandey, A., Singh, L., & Zhang, Z. (2020). Metagenomics for taxonomy profiling: tools and approaches. *Bioengineered*, 11(1), 356–374.
- Bai, Y., Wang, Q., Liao, K., Jian, Z., Zhao, C., & Qu, J. (2018). Fungal community as a bioindicator to reflect anthropogenic activities in a river ecosystem. *Frontiers in Microbiology*, 9, 3152.

- Bellinger, E. G., & Sigeo, D. C. (2015). *Freshwater algae: Identification, enumeration and use as Bioindicators*. Wiley.
- Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., Solimini, A., van de Bund, W., Zampoukas, N., & Hering, D. (2012). Three hundred ways to assess europe's surface waters: An almost complete overview of biological methods to implement the water framework directive. *Ecological Indicators*, 18, 31–41.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer Verlag.
- Bock, C., Jensen, M., Forster, D., Marks, S., Nuy, J., Psenner, R., Beisser, D., & Boenigk, J. (2020). Factors shaping community patterns of protists and bacteria on a European scale. *Environmental Microbiology*, 22(6), 2243–2260.
- Bock, C., Salcher, M., Jensen, M., Pandey, R. V., & Boenigk, J. (2018). Synchrony of eukaryotic and prokaryotic planktonic communities in three seasonally sampled austrian lakes. *Frontiers in Microbiology*, 9, 1290.
- Boenigk, J., Wodniok, S., Bock, C., Beisser, D., Hempel, C., Grossmann, L., Lange, A., & Jensen, M. (2018). Geographic distance and mountain ranges structure freshwater protist communities on a European scale. *Metabarcoding and Metagenomics*, 2, e21519.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Cáceres, M. D., & Legendre, P. (2009). Associations between species and groups of sites: indices and statistical inference. *Ecology*, 90(12), 3566–3574. <https://doi.org/10.1890/08-1823.1>
- Cazelles, K., Araújo, M. B., Mouquet, N., & Gravel, D. (2015). A theory for species co-occurrence in interaction networks. *Theoretical Ecology*, 9(1), 39–48.
- Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, 1(5), e1400253.
- Chaffron, S., Rehrauer, H., Pernthaler, J., & von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7), 947–959.
- Chase, J. M., & Leibold, M. A. (2003). *Ecological Niches – Linking classical and contemporary approaches*. University of Chicago Press.
- Chen, L., Tsui, M. M., Lam, J. C., Hu, C., Wang, Q., Zhou, B., & Lam, P. K. (2019). Variation in microbial community structure in surface seawater from pearl river delta: Discerning the influencing factors. *Science of the Total Environment*, 660, 136–144.
- Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., & Zhao, H. (2013). A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS One*, 8(8), e70837.
- Coclet, C., Garnier, C., Durrieu, G., Omanović, D., D'Onofrio, S., Poupon, C. L., Mullot, J.-U., Briand, J.-F., & Misson, B. (2019). Changes in bacterioplankton communities resulting from direct and indirect interactions with trace metal gradients in an urbanized marine coastal area. *Frontiers in Microbiology*, 10, 257.
- Cordier, T. (2019). Bacterial communities' taxonomic and functional turnovers both accurately predict marine benthic ecological quality status. *Environmental DNA*, 2(2), 175–183. <https://doi.org/10.1002/edn3.55>
- Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, 18(6), 1381–1391.
- Cordier, T., Lanzén, A., Apothéoz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019). Embracing environmental genomics and machine learning for routine biomonitoring. *Trends in Microbiology*, 27(5), 387–397.
- Deltedesco, E., Keiblinger, K. M., Piepho, H.-P., Antonielli, L., Pötsch, E. M., Zechmeister-Boltenstern, S., & Gorfer, M. (2020). Soil microbial community structure and function mainly respond to indirect effects in a multifactorial climate manipulation experiment. *Soil Biology and Biochemistry*, 142, 107704.
- Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z.-I., Knowler, D. J., Lévêque, C., Naiman, R. J., Prieur-Richard, A.-H., Soto, D., Stiassny, M. L. J., & Sullivan, C. A. (2005). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews*, 81(2), 163.
- Dufrêne, M., & Legendre, P. (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs*, 67(3), 345–366.
- Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16), 2194–2200.
- Evans, M. R., Bithell, M., Cornell, S. J., Dall, S. R. X., Díaz, S., Emmott, S., Ermande, B., Grimm, V., Hodgson, D. J., Lewis, S. L., Mace, G. M., Morecroft, M., Moustakas, A., Murphy, E., Newbold, T., Norris, K. J., Petchey, O., Smith, M., Travis, J. M. J., & Benton, T. G. (2013). Predictive systems ecology. *Proceedings of the Royal Society B: Biological Sciences*, 280(1771), 20131452.
- Foissner, W., & Berger, H. (1996). A user-friendly guide to the ciliates (Protozoa, Ciliophora) commonly used by hydrobiologists as bio-indicators in rivers, lakes, and waste waters, with notes on their ecology. *Freshwater Biology*, 35(2), 375–482.
- Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology*, 1–19.
- García-García, N., Tamames, J., Linz, A. M., Pedrós-Alió, C., & Puente-Sánchez, F. (2019). Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. *The ISME Journal*, 13(12), 2969–2983. <https://doi.org/10.1038/s41396-019-0487-8>
- Glasl, B., Bourne, D. G., Frade, P. R., Thomas, T., Schaffelke, B., & Webster, N. S. (2019). Microbial indicators of environmental perturbations in coral reef ecosystems. *Microbiome*, 7, 94.
- Grossmann, L., Beisser, D., Bock, C., Chatzinotas, A., Jensen, M., Preisfeld, A., Psenner, R., Rahmann, S., Wodniok, S., & Boenigk, J. (2016). Trade-off between taxon diversity and functional diversity in European lake ecosystems. *Molecular Ecology*, 25(23), 5876–5888.
- Grossmann, L., Jensen, M., Heider, D., Jost, S., Glücksman, E., Hartikainen, H., Mahamdallie, S. S., Gardner, M., Hoffmann, D., Bass, D., & Boenigk, J. (2016). Protistan community analysis: key findings of a large-scale molecular sampling. *The ISME Journal*, 10(9), 2269–2279.
- Grossmann, L., Jensen, M., Pandey, R. V., Jost, S., Bass, D., Psenner, R., & Boenigk, J. (2016). Molecular investigation of protistan diversity along an elevation transect of alpine lakes. *Aquatic Microbial Ecology*, 78(1), 25–37.
- Guimarães, P. R., Pires, M. M., Jordano, P., Bascompte, J., & Thompson, J. N. (2017). Indirect effects drive coevolution in mutualistic networks. *Nature*, 550(7677), 511–514.
- Han, M., Dsouza, M., Zhou, C., Li, H., Zhang, J., Chen, C., Yao, Q., Zhong, C., Zhou, H., Gilbert, J. A., Wang, Z., & Ning, K. (2019). Agricultural risk factors influence microbial ecology in Honghu Lake. *Genomics, Proteomics & Bioinformatics*, 17(1), 76–90.
- Harris, D. J. (2016). Inferring species interactions from co-occurrence data with Markov networks. *Ecology*, 97(12), 3308–3314.
- Heink, U., & Kowarik, I. (2010). What are indicators? On the definition of indicators in ecology and environmental planning. *Ecological Indicators*, 10(3), 584–593.
- Henson, M. W., Hanssen, J., Spooner, G., Fleming, P., Pukonen, M., Stahr, F., & Thrash, J. C. (2018). Nutrient dynamics and stream order influence microbial community patterns along a 2914 kilometer transect of the Mississippi River. *Limnology and Oceanography*, 63(5), 1837–1855.
- Hering, D., Borja, A., Carstensen, J., Carvalho, L., Elliott, M., Feld, C. K., Heiskanen, A.-S., Johnson, R. K., Moe, J., & Pont, D. (2010). The European Water Framework Directive at the age of 10: A critical

- review of the achievements with recommendations for the future. *Science of the Total Environment*, 408(19), 4007–4019.
- Hermans, S. M., Buckley, H. L., Case, B. S., Curran-Cournane, F., Taylor, M., & Lear, G. (2016). Bacteria as emerging indicators of soil condition. *Applied and Environmental Microbiology*, 83, e02826-16. <https://doi.org/10.1128/AEM.02826-16>
- Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., & Andersson, A. F. (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biology*, 16, 279.
- Isbell, F., Gonzalez, A., Loreau, M., Cowles, J., Díaz, S., Hector, A., Mace, G. M., Wardle, D. A., O'Connor, M. I., Duffy, J. E., Turnbull, L. A., Thompson, P. L., & Larigauderie, A. (2017). Linking the influence and dependence of people on biodiversity across scales. *Nature*, 546(7656), 65–72.
- Jensen, M. (2017). V9\_Clust.R. R-Script for modifying DNA sequence abundance tables: clustering of related sequences (e.g. SSU-ITS1) according to 100% identical sub-sequences. <https://github.com/manfred-uni-essen/V9-cluster>
- Jiang, J.-G., & Shen, Y.-F. (2005). Use of the aquatic protozoa to formulate a community biotic index for an urban water system. *Science of the Total Environment*, 346(1–3), 99–111.
- Johnson, J. L., Fawley, M. W., & Fawley, K. P. (2007). The diversity of Scenedesmus and Desmodesmus (Chlorophyceae) in Itasca State Park, Minnesota, USA. *Phycologia*, 46(2), 214–229.
- Jørgensen, S. E. (2016). *Introduction to Systems Ecology*. CRC Press.
- Karimi, B., Maron, P. A., Boure, N.-C.-P., Bernard, N., Gilbert, D., & Ranjard, L. (2017). Microbial diversity and ecological networks as indicators of environmental quality. *Environmental Chemistry Letters*, 15(2), 265–281.
- Karlsson, J., Jonsson, A., & Jansson, M. (2005). Productivity of high-latitude lakes: climate effect inferred from altitude gradient. *Global Change Biology*, 11(5), 710–715.
- Kegler, H. F., Hassenrück, C., Kegler, P., Jennerjahn, T. C., Lukman, M., Jompa, J., & Gärdes, A. (2018). Small tropical islands with dense human population: differences in water quality of near-shore waters are associated with distinct bacterial communities. *PeerJ*, 6, e4555.
- Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.-M., Humbert, J.-F., & Bouchez, A. (2014). A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, 33(1), 349–363.
- Kiersztyn, B., Chróst, R., Kaliński, T., Siuda, W., Bukowska, A., Kowalczyk, G., & Grabowska, K. (2019). Structural and functional microbial diversity along a eutrophication gradient of interconnected lakes undergoing anthropopressure. *Scientific Reports*, 9, 11144.
- Krikorian, N. (1979). The Volterra model for three species predator-prey systems: Boundedness and stability. *Journal of Mathematical Biology*, 7(2), 117–132.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1–26.
- Lange, A., Jost, S., Heider, D., Bock, C., Budeus, B., Schilling, E., Strittmatter, A., Boenigk, J., & Hoffmann, D. (2015). AmpliconDuo: A split-sample filtering protocol for high-throughput amplicon sequencing of microbial communities. *PLoS One*, 10(11), e0141590.
- Lee, S., Basu, S., Tyler, C. W., & Wei, I. W. (2004). Ciliate populations as bio-indicators at deer island treatment plant. *Advances in Environmental Research*, 8(3–4), 371–378.
- Legendre, P., & Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2), 271–280.
- Logares, R., Tesson, S. V., Canbäck, B., Pontarp, M., Hedlund, K., & Rengefors, K. (2018). Contrasting prevalence of selection and drift in the community structuring of bacteria and microbial eukaryotes. *Environmental Microbiology*, 20(6), 2231–2240. <https://doi.org/10.1111/1462-2920.14265>
- Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, 2, e593.
- Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, 13(1), 31.
- Massana, R., & Logares, R. (2012). Eukaryotic versus prokaryotic marine picoplankton ecology. *Environmental Microbiology*, 15(5), 1254–1261.
- Merkley, M., Rader, R. B., McArthur, J. V., & Eggett, D. (2004). Bacteria as bioindicators in wetlands: Bioassessment in the Bonneville Basin of Utah, USA. *Wetlands*, 24(3), 600–607.
- Mikhailov, I. S., Zakharova, Y. R., Bukin, Y. S., Galachyants, Y. P., Petrova, D. P., Sakirko, M. V., & Likhoshway, Y. V. (2018). Co-occurrence networks among bacteria and microbial eukaryotes of Lake Baikal during a spring phytoplankton bloom. *Microbial Ecology*, 77, 96–109.
- Miller, T. E., & Travis, J. (1996). The evolutionary role of indirect effects in communities. *Ecology*, 77(5), 1329–1335.
- Nuy, J. K., Hoetzing, M., Hahn, M. W., Beisser, D., & Boenigk, J. (2020). Ecological differentiation in two major freshwater bacterial taxa along environmental gradients. *Frontiers in Microbiology*, 11, 154.
- Otwell, A. E., de Lomana, A. L. G., Gibbons, S. M., Orellana, M. V., & Baliga, N. S. (2018). Systems biology approaches towards predictive microbial ecology. *Environmental Microbiology*, 20(12), 4197–4209.
- Panwar, P., Allen, M. A., Williams, T. J., Hancock, A. M., Brazendale, S., Bevington, J., Roux, S., Páez-Espino, D., Nayfach, S., Berg, M., Schulz, F., Chen, I. A., Huntemann, M., Shapiro, N., Kyrpides, N. C., Woyke, T., Elie-Fadrosh, E. A., & Cavicchioli, R. (2020). Influence of the polar light cycle on seasonal dynamics of an antarctic lake microbial community. *Microbiome*, 8, 116.
- Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11), 1533–1542. <https://doi.org/10.1038/s41564-017-0012-7>
- Pershina, E., Valkonen, J., Kurki, P., Ivanova, E., Chirak, E., Korvigo, I., Provorov, N., & Andronov, E. (2015). Comparative analysis of prokaryotic communities associated with organic and conventional farming systems. *PLoS One*, 10(12), e0145072.
- Plassart, P., Prévost-Bouré, N. C., Uroz, S., Dequiedt, S., Stone, D., Creamer, R., Griffiths, R. I., Bailey, M. J., Ranjard, L., & Lemanceau, P. (2019). Soil parameters, land use, and geographical distance drive soil bacterial communities along a European transect. *Scientific Reports*, 9, 605.
- Proulx, S. R., Promislow, D. E., & Phillips, P. C. (2005). Network thinking in ecology and evolution. *Trends in Ecology & Evolution*, 20(6), 345–353.
- Puranik, S., Pal, R. R., More, R. P., & Purohit, H. J. (2016). Metagenomic approach to characterize soil microbial diversity of Phumdi at Loktak Lake. *Water Science and Technology*, 74(9), 2075–2086.
- Reis, M. P., Suhadolnik, M. L. S., Dias, M. F., Ávila, M. P., Motta, A. M., Barbosa, F. A., & Nascimento, A. M. (2020). Characterizing a riverine microbiome impacted by extreme disturbance caused by a mining sludge tsunami. *Chemosphere*, 253, 126584.
- Röttgers, L., & Faust, K. (2018a). Can we predict keystones? *Nature Reviews Microbiology*, 17(3), 193.
- Röttgers, L., & Faust, K. (2018b). From hairballs to hypotheses – biological insights from microbial networks. *FEMS Microbiology Reviews*, 42(6), 761–780.
- Schlegel, I., Koschel, R., & Krienitz, L. (1998). On the occurrence of *Phacotus lenticularis* (Chlorophyta) in lakes of different trophic state. *Hydrobiologia*, 369(370), 353–361.
- Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6), 863–864.

- Sharuddin, S. S., Ramli, N., Hassan, M. A., Mustapha, N. A., Amran, A., Mohd-Nor, D., Sakai, K., Tashiro, Y., Shirai, Y., & Maeda, T. (2017). Bacterial community shift revealed Chromatiaceae and Alcaligenaceae as potential bioindicators in the receiving river due to palm oil mill effluent final discharge. *Ecological Indicators*, 82, 526–529.
- Sperlea, T., Füser, S., Boenigk, J., & Heider, D. (2018). SEDE-GPS: socio-economic data enrichment based on GPS information. *BMC Bioinformatics*, 19, 440.
- Steffen, W., Persson, A., Deutsch, L., Zalasiewicz, J., Williams, M., Richardson, K., Crumley, C., Crutzen, P., Folke, C., Gordon, L., Molina, M., Ramanathan, V., Rockström, J., Scheffer, M., Schellnhuber, H. J., & Svedin, U. (2011). The Anthropocene: From global change to planetary stewardship. *Ambio*, 40(7), 739–761.
- Tan, B., Ng, C., Nshimiyimana, J. P., Loh, L. L., Gin, K.-Y.-H., & Thompson, J. R. (2015). Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges and future opportunities. *Frontiers in Microbiology*, 6, 1027.
- Thessen, A. E., Patterson, D. J., & Murray, S. A. (2012). The taxonomic significance of species that have only been observed once: The genus *Gymnodinium* (Dinoflagellata) as an example. *PLoS One*, 7(8), e44015.
- Tilman, D., Clark, M., Williams, D. R., Kimmel, K., Polasky, S., & Packer, C. (2017). Future threats to biodiversity and pathways to their prevention. *Nature*, 546(7656), 73–81.
- Ulanowicz, R. E. (2001). Information theory in ecology. *Computers & Chemistry*, 25(4), 93–399.
- Vishnivetskaya, T. A., Almatari, A. L., Spirina, E. V., Wu, X., Williams, D. E., Pfiffner, S. M., & Rivkina, E. M. (2020). Insights into community of photosynthetic microorganisms from permafrost. *FEMS Microbiology Ecology*, 96(12), fiae229.
- Washburne, A. D., Silverman, J. D., Leff, J. W., Bennett, D. J., Darcy, J. L., Mukherjee, S., Fierer, N., & David, L. A. (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, 5, e2969.
- Webster, N. S., Wagner, M., & Negri, A. P. (2018). Microbial conservation in the Anthropocene. *Environmental Microbiology*, 20(6), 1925–1928.
- Wei, J., Gao, J., Wang, N., Liu, Y., Wang, Y., Bai, Z., Zhuang, X., & Zhuang, G. (2019). Differences in soil microbial response to anthropogenic disturbances in Sanjiang and Momoge Wetlands, China. *FEMS Microbiology Ecology*, 95(8), fiz110. <https://doi.org/10.1093/femsec/fiz110>
- Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5, 27.
- Welzel, M., Lange, A., Heider, D., Schwarz, M., Freisleben, B., Jensen, M., Boenigk, J., & Beisser, D. (2020). Natrix: a snakemake-based workflow for processing clustering and taxonomically assigning amplicon sequencing reads. *BMC Bioinformatics*, 21, 526.
- Williams, M., Zalasiewicz, J., Haff, P., Schwägerl, C., Barnosky, A. D., & Ellis, E. C. (2015). The Anthropocene biosphere. *The Anthropocene Review*, 2(3), 196–219. <https://doi.org/10.1177/2053019615591020>
- Williamson, C. E., Dodds, W., Kratz, T. K., & Palmer, M. A. (2008). Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. *Frontiers in Ecology and the Environment*, 6(5), 247–254.
- World Wildlife Fund (2018). *Living planet report*. WWF.
- Yang, Y., Li, S., Gao, Y., Chen, Y., & Zhan, A. (2019). Environment-driven geographical distribution of bacterial communities and identification of indicator taxa in Songhua River. *Ecological Indicators*, 101, 62–70.
- Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In T. Fawcett, & N. Mishra (Eds.). *Proceedings of the Twentieth International Conference on Machine Learning* (pp. 856–863).
- Zeng, Y., Baumbach, J., Barbosa, E. G. V., Azevedo, V., Zhang, C., & Koblížek, M. (2016). Metagenomic evidence for the presence of phototrophic Gemmatimonadetes bacteria in diverse environments. *Environmental Microbiology Reports*, 8(1), 139–149.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Sperlea T, Kreuder N, Beisser D, Hattab G, Boenigk J, Heider D. Quantification of the covariation of lake microbiomes and environmental variables using a machine learning-based framework. *Mol Ecol*. 2021;00:1–14. <https://doi.org/10.1111/mec.15872>

### 2.3 PUBLICATION III

Theodor Sperlea, Jan Philip Schenk, Hagen Dreßler, Daniela Beisser, Georges Hattab, Jens Boenigk, and Dominik Heider (2021). Covariation between the european lake microbiome and surrounding land cover and bioindicator-based insights into the microbiome structure. *In Review at ISME Communications*

**Contributions:** T. Sperlea designed and performed all analyses, JPS provided the OSM dataset, HD helped with the network analyses, DB and JB provided the sequencing datasets, JB, GH, and DH supervised the study. All authors discussed the results and wrote and revised the manuscript.

\* \* \*

Lake ecosystems are impacted by the composition of the landscapes surrounding them. Changes in land cover, often brought about by anthropogenic processes, are of special interest for lake ecology, as they impact the physico-chemical makeup of the lake's water and, with it, the lake's ecology. To understand the relationship between land cover and the composition of the lake microbiome, we combined tools and methods from the previous publications presented here.

We extracted land cover data from both the Open Street Map and the CORINE Land Cover web services. Based on relative areas of distinct land cover categories around the lakes, we attempted to predict the microbial biodiversity as represented by different alpha-diversity metrics found in the lakes but found only low levels of predictability. In contrast, by applying the covariation framework, we identified separate land cover categories that the lake microbiome does covary with. The results presented in this paper underline that aggregating different land cover categories (to super-categories) or organisms (to biodiversity metrics) can obfuscate the relationship that exists between land cover and microbiome.

By extracting bioindicators for the land cover categories, we identified multi-target bioindicators and environmental drivers of microbiome composition, such as the altitude of the lake. Furthermore, after merging the bioindicators identified here with those identified in publication II for physico-chemical parameters, we derived two new data abstractions. One of these indicates which microbes respond to environmental changes in a similar fashion and the other points to environmental parameters that the microbiome responds to comparably. Both result in qualitative insights into the functioning of the European lake microbiome.

\* \* \*

Der ökologische Zustand von Seen wird von umgebenden Landschaften beeinflusst. Veränderungen der Landbedeckung, die oft durch anthropogene Prozesse hervorgerufen werden, sind für die Ökologie von Seen von besonderem Interesse, da sie die physikalisch-chemische Zusammensetzung des Seewassers und damit die Ökologie des Sees beeinflussen. Um die Beziehung zwischen Landbedeckung und der Zusammensetzung des Mikrobioms des Sees zu erforschen, kombinierten wir Werkzeuge und Methoden aus den hier vorgestellten früheren Publikationen.

Wir extrahierten Daten zur Landbedeckung um die beprobten Seen aus den Webdiensten OpenStreetMap und CORINE Land Cover und versuchten, anhand der relativen Flächen der verschiedenen Landbedeckungskategorien, die mikrobielle Biodiversität in den Seeproben vorherzusagen. Wir fanden jedoch nur ein geringes Maß an Vorhersagbarkeit der verschiedenen Alpha-Diversitätsmetriken der Seemikrobiomen. Im Gegensatz dazu konnten wir durch die Anwendung des *covariation framework* einzelne Landbedeckungskategorien identifizieren, mit denen das Mikrobiom der Seen signifikant kovariiert. Die in dieser Arbeit vorgestellten Ergebnisse unterstreichen, dass die Aggregation verschiedener Landbedeckungskategorien (zu Superkategorien) oder Organismen (zu Biodiversitätsmetriken) die Beziehung zwischen Landbedeckung und Mikrobiom verschleiern kann.

Durch die Extraktion von Bioindikatoren für die Landbedeckungskategorien identifizierten wir Multi-Target-Bioindikatoren und Umwelttreiber der Mikrobiom-Zusammensetzung, wie z. B. die Höhenlage des Sees. Darüber hinaus haben wir nach der Zusammenführung der hier identifizierten Bioindikatoren mit denen, die in Publikation II für physikalisch-chemische Parameter identifiziert wurden, zwei neue Datenabstraktionen abgeleitet. Eine davon zeigt an, welche Mikroben auf Umweltveränderungen in ähnlicher Weise reagieren und die andere weist auf Umweltparameter hin, auf die das Mikrobiom vergleichbar reagiert. Beide führen zu qualitativen Erkenntnissen über die Funktionsweise des europäischen Seenmikrobioms.



# Covariation Between the European Lake Microbiome and Surrounding Land Cover and Bioindicator-Based Insights Into the Microbiome Structure

Theodor Sperlea<sup>1</sup>    Jan Philip Schenk<sup>1</sup>    Hagen Dreßler<sup>1</sup>    Daniela Beisser<sup>2</sup>  
Georges Hattab<sup>1</sup>    Jens Boenigk<sup>2</sup>    Dominik Heider<sup>1\*</sup>

<sup>1</sup>Faculty of Mathematics and Computer Science, University of Marburg,  
Hans-Meerwein-Str. 6, D-35032 Marburg (Lahn), Germany

<sup>2</sup>Department of Biodiversity, Center for Water and Environmental Research, University of  
Duisburg-Essen, D-45141 Essen, Germany

\*Corresponding author: Dominik Heider (dominik.heider@uni-marburg.de)

Running Title: Covariation Between Lake Microbiome and Land Cover

## Abstract

Microbes such as bacteria, archaea, and protists are essential for element cycling and ecosystem functioning, but many questions central to the understanding of microbial ecology are still open. Here, we analyze the relationship between lake microbiomes and the land cover surrounding the lakes. By applying machine learning methods, we quantify the covariance between land cover categories and the microbial community composition recorded in the largest amplicon sequencing dataset of European lakes available to date. We identify microbial bioindicators for these land cover categories. Combining land cover and physico-chemical bioindicators identified from the same amplicon sequencing dataset, we develop two novel similarity metrics that facilitate insights into the ecology of the lake microbiome. We show that the bioindicator network, i.e., the graph linking OTUs indicative of the same environmental parameters, corresponds to microbial co-occurrence patterns. Furthermore, we determine environmental parameters the microbiome responds to in a similar manner. Taken together, this paper presents a set of novel

methods that facilitate the study of environmental microbiomes as complex systems and apply them to the European lake microbiome.

## Introduction

Ecosystems are governed by processes at very different scales, ranging from individual metabolic reactions to changes in land use at the landscape level (1, 2, 3). Because the processes that link ecological scales are hard to study in well-controlled experimental settings, we poorly understand how land use shapes the composition of environmental microbiomes. In contrast, observational studies could, in principle, provide these insights but come with considerable methodological and theoretical obstacles.

Lakes accumulate water from their catchment, and with it, nutrients, stressors, and pollutants, making them sentinels of environmental change of the landscape the lakes are part of (4, 5). Microbes, in turn, play an essential role in the functioning and the stability of ecosystems and have been called both “ubiquitous janitors of the Earth” and “first responders” to environmental change (6, 7, 8). Indeed, microorganisms have been and are being used as bioindicators for ecosystem integrity in monitoring schemes (9, 10). However, only the relatively recent advent of next-generation sequencing and environmental metabarcoding expanded the pool of potential bioindicators from visually distinct organisms to all microorganisms (11, 12).

One of the most widely used methods for the identification of bioindicators is the indicator value (IndVal) method (13, 14, 15). Given multiple groups of sites, defined by, e.g., high, medium, and low values for a parameter of interest, it determines sets of organisms that are indicative for each of the groups of sites. The relationship between bioindicator and the indicated environmental parameter is apparent (and therefore of use for biomonitoring schemes) but not necessarily direct or causal (16, 17). This follows from the ecosystem being a complex system, i.e., centrally defined by non-linear interactions of the biotic and abiotic factors in it (18, 19). The response of an organism to an environmental signal can be modulated by the presence and abundance of the other organisms in the ecosystem (20, 21). Thus, organisms that strongly interact, e.g., symbiotically, will be indicative of the same parameters. Similarly, the interaction between the environmental parameter in question and the bioindicator might be relayed through other environmental parameters. For example, land cover only influences lake microbes indirectly, through, e.g., physico-chemical water parameters. Furthermore, land cover areas are not independent because an increase in one necessarily leads to decreases in others (22).

In a prior publication, we quantified the covariation between a set of physico-chemical parameters and



the lake microbiome, i.e., the prevalence of all microorganisms in a collection of environmental samples, reported in an OTU table derived from amplicon sequencing (23). At its core, the covariation framework works similar to the regular machine learning workflow (for details, see Methods), but with a slight twist on its interpretation: In a supervised regression task, we train a model to approximate the relationship of a set of input variables  $X$  to the target variable  $y$ . Assuming that the model can capture most of the patterns present in  $X$  relevant for the prediction of  $y$ , we can interpret the model’s prediction  $\hat{y}$  as a non-linear projection of  $X$  into the space of  $y$ . In this context, the coefficient of determination,  $R^2$ , calculated between  $y$  and  $\hat{y}$  can intuitively be interpreted as the covariation between the microbiome as a whole and the environmental parameter in question. The covariation framework, thus, circumvents both the obstacles described above: First, it uses machine learning methods that can model non-linear dependencies in non-independent data, like Random Forests (24), to handle the interdependencies between microbial species in the dataset. Second, it explicitly avoids any association with direct interaction, correlation, let alone causal relationships between the microbiome and the environmental parameters in question. Furthermore, as it employs the IndVal method as the feature selection method, the covariation framework identifies microbial bioindicators and quantifies their predictive power.

In this paper, we apply the covariation framework to study the relationship between the land cover, i.e., the type and usage of the Earth’s surface, that surrounds lakes and these lakes’ microbiome. To this end, we assessed land cover data from the OpenStreetMap (OSM) project as well as the CORINE Land Cover (CLC) dataset from the Copernicus Land Monitoring Service (25, 26). The former of the two data sources provides an open, community-driven, and, thus, rather detailed but potentially incomplete land cover categorization. In contrast, the latter dataset is based on high-resolution satellite imagery and contains a hierarchical categorization of land cover in 44 classes. After reporting on the microbial land cover covariation and presenting microbial land cover bioindicators, we combine these with the bioindicators for physico-chemical parameters identified in (23) for further analyses. From this collection of bioindicators, we first derive multi-target bioindicators, i.e., species with a high significance for ecosystem functioning that might act as keystone species. Second, we derive a similarity matrix of bioindicators based on the parameters they are indicative for and show that this corresponds to their co-occurrence. Finally, we use the lists of species indicative for the environmental parameters as a metric of similarity of the microbiome’s response to that parameter and present a “response map” of the microbiome based on bioindicator analysis.

## Materials and methods

### Amplicon sequencing

Sampling was part of a pan-European study conducted in August 2012 (eukaryotic sequences are published in (27); NCBI Bioproject PRJNA414052, prokaryotic sequences are published and described in (28) and (29); NCBI Bioproject PRJNA559862). Methods for data collection, extraction, sequencing, and amplicon processing are described in detail in these studies (27, 28, 29) and will be briefly outlined below.

To analyze bacterial and protistan freshwater communities on a large scale, 280 lakes were sampled throughout Europe. Sampling details and information on measured physico-chemical and geographical parameters can be found in (27). For DNA analyses, filtered water samples were air-dried and frozen in liquid nitrogen. Genomic DNA was extracted using the my-Budget DNA Mini Kit (Bio-Budget Technologies GmbH, Krefeld, Germany) with modifications after (27). Amplicon sequencing targeted the V2-V3 region of the 16S rRNA gene for bacteria, the V9 region of the 18S, and the ITS1 gene for eukaryotes. Samples were commercially sequenced (Fasteris, Geneva, Switzerland) on an Illumina HiSeq 2500 sequencer generating 300 bp long paired-end reads. Adapter removal, quality trimming, and demultiplexing were performed by the sequencing company.

Sequence processing was performed using a provisional version of the Natrix pipeline (30). If not stated otherwise, all software versions and parameters were used as described in (30). The main steps included quality checks using FASTQC (31) and PRINSEQ (32), assembly of paired-end reads with PANDASeq (33) and dereplication and chimera removal using UCHIME (usearch v7.0.1090 with default parameters) (34). AmpliconDuo (35) was used to discard sequences that were not found in both technical replicates. The remaining sequences were clustered using SWARM (36) and further aggregated to identical V9 sequences. This aggregation served as the basis for the OTU tables. The taxonomic assignment of the eukaryotic sequences was performed by a BLAST search (37) against the NCBI nt database and for the prokaryotic sequences against SILVA SSURf 132 (38). For all downward analyses, we combined the prokaryotic and eukaryotic OTU tables.

### Land cover data

Two different land cover datasets were used in this study. For both, we accessed data for the year 2012 because this was also the year the lake samples were collected. The CLC dataset was downloaded from the official website of Copernicus Earth Observation program (CLC 2012, v.2020\_20u1, 100m raster GeoTiff)(26). The

relative areas of the land cover classes were extracted from the dataset for circular areas around the sampling points with different radii using QGIS 3.16 (39). Areas were aggregated to higher-level land cover classes according to the hierarchical CLC class model.

OSM land cover data was extracted from the OSM planet file from September 2012 archived at archive.org. This file was loaded in a PostgreSQL database and queried using a routine adapted from SEDE-GPS (40) to retrieve the map tiles surrounding the sampling position, to fuse these, and to extract a circular area of a given radius. Map tiles were rendered using the default mapnik map style, which was adjusted to (i) merge pixels of land use sub-categories with the respective main category (such as “tertiary road” with “road”) and (ii) remove signs, labels, and point of interest markers. The pixel-areas summarised per unique category of the resulting image were read out and translated back to meters.

Outlier land cover values in all subsets (concerning both the radius as well as the land cover category) of these two datasets were detected using the function *boxplot.stats* in R 4.0.3 for the different radii and land cover categories, separately. Samples containing an outlier or a value of zero for a given land cover category at the given radius were discarded for the analysis of the respective land cover category and radius.

Additional physico-chemical parameters were taken from (23).

## Prediction of biodiversity from land cover

To calculate biodiversity metrics, the OTU table was rarefied using the *rrarefy* function from the R package **vegan** (v2.5-6, ref. 41). Biodiversity metrics were calculated from rarefied OTU tables using the *diversity* (Shannon index, Simpson diversity, inverse Simpson diversity), and *renyi* (Renyi entropy) functions from the R package **vegan**, except for species richness, which is the total number of OTUs present, and Pielou’s evenness, which was calculated by dividing the sample’s Shannon index by the log of the sample’s richness (42, 43). Renyi entropy metrics were calculated for  $\alpha = \{0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, \infty\}$ , as different values for this parameter drastically change the metrics sensitivity to relative species abundance (42). For the prediction of biodiversity metrics based on land cover categories, Random Forests from the R package **caret** (version 6.0.86, ref. 44) were trained using 10-fold cross-validation without further feature selection.

## Covariation framework

Covariation between the lake microbiomes and land cover areas was quantified using the covariation framework presented in ref. 23. Methodologically, the framework is a straightforward machine learning approach, training a machine learning model to predict the values of an environmental parameter after feature selection.

However, the model’s prediction is interpreted as the projection of the microbial community composition to the space of the target variable while leveraging the model’s potential to model non-linear interdependencies in the microbiome. This way, the coefficient of determination,  $R^2$ , can be interpreted as a measure of covariation between the microbiome as a whole and the target variable. As feature selection method, the *multipatt* function (with the parameter “indval”) from the *indicspecies* R package (v1.7.9, ref. 13) was used after Hellinger transformation (45) of the OTU counts to identify bioindicator OTUs for tertiles of the respective land cover category. Random Forest models from the R package *caret* (version 6.0.86, ref. 44) were trained in a 10-fold cross-validation scheme with the OTU tables as independent and the relative area of a single land cover category as dependent variables, with both being centered and log-ratio transformed. Because of statistical limitations of the Random Forest model, some combinations of area size and land cover category resulted could not be used for model training.

Confidence intervals for the model evaluations were estimated based on resampling of predicted and measured dependent variable pairs with replacement with thousand repetitions. Statistical significance of relevant models was asserted by comparing the  $R^2$  value with results gathered by thousand repetitions of training models with the same hyper-parameter setting on resampled biodiversity data in a Student’s t-test as implemented in the *t.test* function in R 4.0.3.

## Bioindicator analysis

Bioindicator OTUs for land cover categories were identified using the indicator species method as implemented in the *multipatt* function in the *indicspecies* R package (v1.7.9) (13, 15). A significance level of  $\alpha = 0.05$  was applied after Benjamini-Hochberg correction for the total number of land cover categories analyzed in this study.

The similarity of two environmental parameters was calculated in terms of the microbiome’s response to changes in them by calculating the Jaccard similarity between the lists of OTUs indicative of the two parameters. The resulting similarity matrix was visualized as a force embedded network using the function *qgraph* from the package *qgraph* (v1.6.9, ref. 46). Furthermore, a dissimilarity matrix was derived from the similarity matrix by inversion after the addition of a random number in the order of  $10^{-8}$  in order to avoid the division by zero. This dissimilarity matrix was visualized as a dendrogram using *upgma* and *ggdendrogram* from the packages *phangorn* (v2.5.5., ref. 47) and *ggdendro* (v0.1.22, ref. 48), respectively. For the ordination of the environmental parameters, the *metaMDS* function from the R package *vegan* was used.

For the bioindicator network, an edge was created between all pairs of bioindicator OTUs that are indicative of at least one common environmental parameter. Null-hypothesis networks were created the same way based on resampled indicator lists; for these, each lake parameter is assigned the same number of randomly selected OTUs as pseudo-indicators in such a way that the distribution of cardinalities is the same as that of the real OTUs. Node properties (degree, closeness centrality, eigenvector centrality, page rank, and authority score) were calculated using the `igraph` package in R 4.0.3 (v1.2.6, ref. 49).

## Network inference methods

In this study, we apply several methods for the inference of network structures from OTU tables. Most of these employ similarity measures as edge weights calculated between all pairs of OTUs, which act as nodes in the network. Simple co-occurrence, checkerboard score (50, 51), Bray-Curtis similarity (52), Kullback-Leibler divergence (53), Pearson and Spearman correlation were used as similarity metrics. The co-occurrence metric was defined as the number of samples in which the two OTUs in question had non-zero occurrence. Pearson and Spearman correlations were calculated using the `cor` function in R and results below the significance level  $\alpha = 0.05$  were discarded. For the calculation of the Kullback-Leibler divergence, zeroes in the OTU table were replaced by  $10^{-8}$  to avoid infinities created by the logarithm. Additionally, the method SparCC (54) was used as implemented in the `sparcc` function from the `SpiecEasi` package (v1.1.0, ref. 55) with default parameters. For all networks, OTUs that have zero counts for 25 or more sampling sites were excluded from the analysis to avoid statistical artifacts that are based on the rarity of the OTUs in question (56).

All figures were generated using the R packages `ggplot2` (v.3.3.2, ref. 57), unless otherwise noted, and following the guidelines laid out in ref. 58.

## Results

### Low predictability of microbial biodiversity from land cover

A straightforward way of determining whether land cover changes impact lake microbiomes is to assess whether the distribution of land cover types surrounding the lake is predictive of the lake's microbial biodiversity. To this end, we extracted the relative area covered by different land cover categories in circular areas around the sampling sites of the European lake dataset from both the CORINE land cover (CLC) dataset as well as the OpenStreetMap (OSM) project (see Methods). These two datasets differ in the way they were generated and their categorization of land cover. While the former is derived from satellite data, the latter

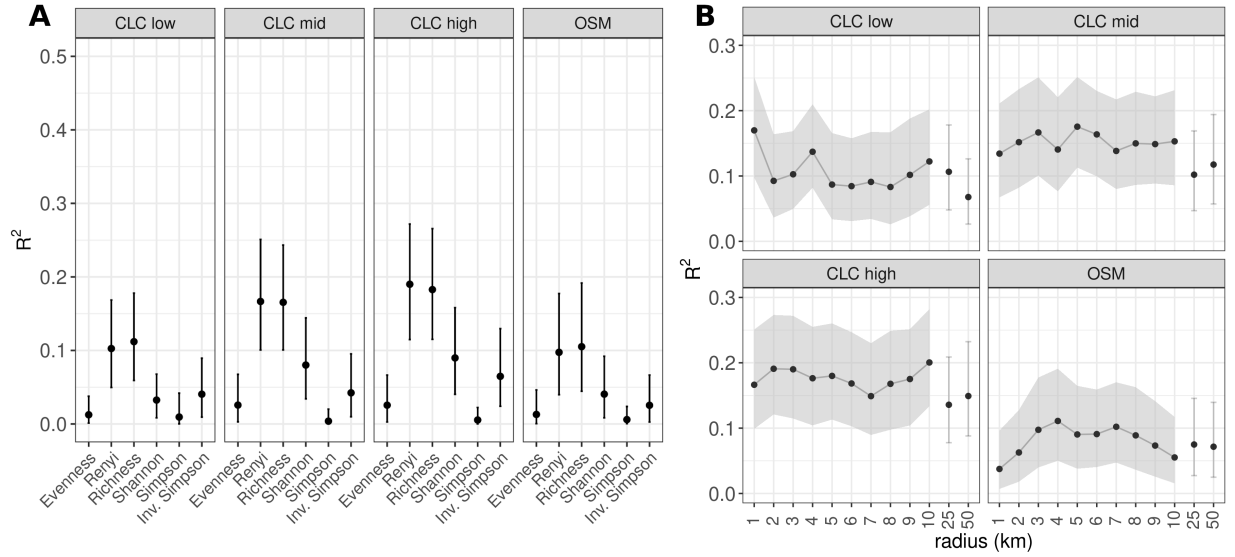


Figure 1: Evaluation of Random Forest models trained to predict biodiversity metrics from land cover. **A.** Results for the land cover in a radius of 3 km around the sampling site and all biodiversity metrics surveyed in this study. Lines represent confidence intervals estimated based on resampling (see methods). For results of other radii, and results for Renyi entropy with other values for  $\alpha$ , see supplementary figures 1 and 2. **B.** Results for the full range of radii as well as 25 and 50 km for Renyi entropy with  $\alpha = 0.5$ . Grey areas and error bars represent confidence intervals; results for 25 and 50 km are visually separated to underline that these radii are not in the range of the other radii.

is annotated in a community-driven manner, based on landscape features observed “on the ground”. To distinguish between effects present at shorter or longer geographic ranges, we extracted and analyzed areas surrounding the sampling points within radii ranging from 1 km to 10 km, in steps of 1 km, as well as 25 km and 50 km. We then assessed the degree to which the relative sub-areas of the land cover categories contained in the extracted areas can be used to predict a set of biodiversity metrics calculated for the microbial communities of the sampled lakes using Random Forest models.

Our results suggest that there is, at best, a marginal predictive relationship between land cover and microbial biodiversity, as no combination of radius, biodiversity metric, and dataset result in  $R^2 > 0.2$  (see figure 1A, supplementary figure 1 and 2). For all radii studied here, the lake microbiome’s Renyi entropy is most predictable from land cover, followed by species richness. In contrast to our expectations, we found no significant difference between  $R^2$  values obtained for the same biodiversity metric at different radii (see figure 1B), indicating that the results presented here are most likely due to statistical artifacts rather than processes that shape microbial biodiversity based on surrounding land cover. In general, the land cover data collected from the OSM dataset is less predictive for microbial biodiversity than the CLC datasets (see figures 1A and B). Therefore, we focus our further analysis on the CLC dataset. Taken together, these results suggest that

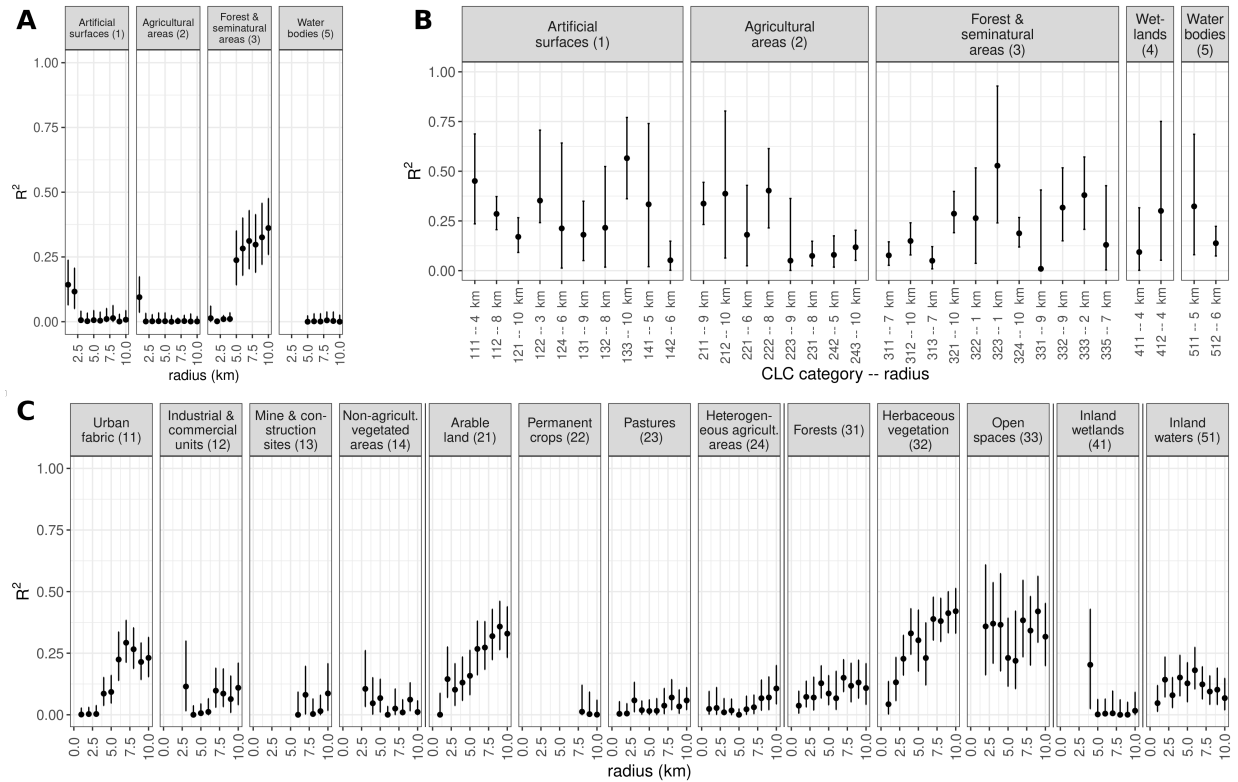


Figure 2: Covariation between land cover surrounding lakes and the lake’s microbiome. Numbers in brackets (**A** and **C**) and in x-axis labels (**B**) refer to the CLC category number code (see table 1). For ease of display, full-length labels of the CLC categories have been shortened in some cases. Vertical lines represent confidence intervals estimated from resampling (see methods) and dots represent covariation as obtained in model evaluation. **A.** Results for all high-level land cover categories in the CLC dataset. **B.** Results for low-level land cover categories; for each land cover category, only the results for the radius with the highest  $R^2$  are shown. **C.** Results for all mid-level land cover categories in the CLC dataset. Vertical lines between facets separate groups of categories that are subcategories of the categories in **A**. For all results, see supplementary table 1.

if land cover has a structuring effect on lake microbiomes, biodiversity metrics do not reflect this effect.

## Microbial community structure covaries with specific land cover categories

To assess whether there is a relationship between land cover surrounding the lakes and the lake microbial community composition at the OTU level, we applied the covariation framework to the land cover categories present in the CLC datasets. Higher  $R^2$  values indicate a higher degree of covariation between the microbiome as a whole and the target parameter, i.e., the relative area of the land use category in question. On the highest level of CLC category hierarchy, we observe covariation of  $R^2 > 0.05$  for “artificial surfaces (1)” and “agricultural areas (2)” at very low radii as well as increasing covariation for “forest and semi-natural areas

(3)” with increasing radii (see figure 2A). In contrast, the anthropogenic effects that are expected with built surfaces and agriculturally used land (59, 60) act on rather short ranges.

The covariation observed between the lake microbiome and land cover categories from the middle level of the CLC hierarchy paints a more nuanced picture (see figure 2C). For example, we observe increasing covariation with the lake microbiomes at increasing radii for the land cover categories “arable land (21)” and “scrub and/or herbaceous vegetation associations (32)”. In contrast, for “urban fabric (11)” and “inland waters (51)”, we observe a peak in covariation at radii of 7 km and 6 km, respectively, with lower  $R^2$  for the other radii. For “forests (31)” and “open spaces with little or no vegetation (33)”, the covariation for different radii stays within the respective confidence intervals of the covariation at the 1 km radius. Notably, the covariation between the microbiome and sub-categories of a CLC category can deviate strongly from the covariation between the microbiome and the respective super-category. The same goes for covariations observed at the lowest level of the CLC category hierarchy (see figure 2B). Taken together, these results show that a broad array of land cover categories have an impact on the lake’s microbial community composition at the OTU level. Furthermore, different land cover categories show the highest covariation with the microbial community composition at different radii, suggesting that other mechanisms are at play for different land cover categories.

To identify general spatial trends, we separately calculated the mean covariation of all land cover categories for each radius and land cover hierarchy level. For all but one radius-hierarchy level combination, the average  $R^2$  value is below 0.15 (see supplementary figure 3). Throughout all combinations, the relative standard deviation is close to or higher than 100%. This result shows that there are neither general spatial trends nor a generally higher covariation at lower levels of the CLC hierarchy.

## Microbial lake bioindicators for surrounding land cover categories

Using the indicator value method that is part of the covariation framework, we identified 2,354 OTUs that act as bioindicators for the land cover categories in a total of 4,453 indicator-parameter pairs (for a complete list, see supplementary table 2). Among the 27 land cover categories studied in this paper, for “scrub and/or herbaceous vegetation associations (32)” and “forest and semi-natural areas (3)” we identify the highest number of indicator OTUs with 1056 and 703 OTUs, respectively (see table 1). Most of the indicator OTUs are Bacteria (87%) from the phyla Proteobacteria (29%), Bacteroidetes (28%), or Cyanobacteria (15%), and from the classes Flavobacteriia (16%) or Alphaproteobacteria (13%). Furthermore, most of the OTUs obtained are indicative of more than one land cover parameter (fig. 3A). All OTUs indicative of more than



seven land cover parameters are bacteria (see tables 2 and 3). This supports the notion that bacteria are more sensitive to environmental changes or respond to environmental signals in a different manner than microbial eukaryotes (29, 61).

In a previous paper, we identified bioindicator OTUs for physico-chemical parameters while working with the same amplicon sequencing dataset as analyzed here (23). Comparing the results of this analysis with those in the prior publication, we observed that almost all bioindicators indicative for at least eight land cover categories are also indicative of the lake's altitude (see table 2). This underlines the central role the geographic location of a lake plays in the lake's ecology (see discussion) (29, 62).

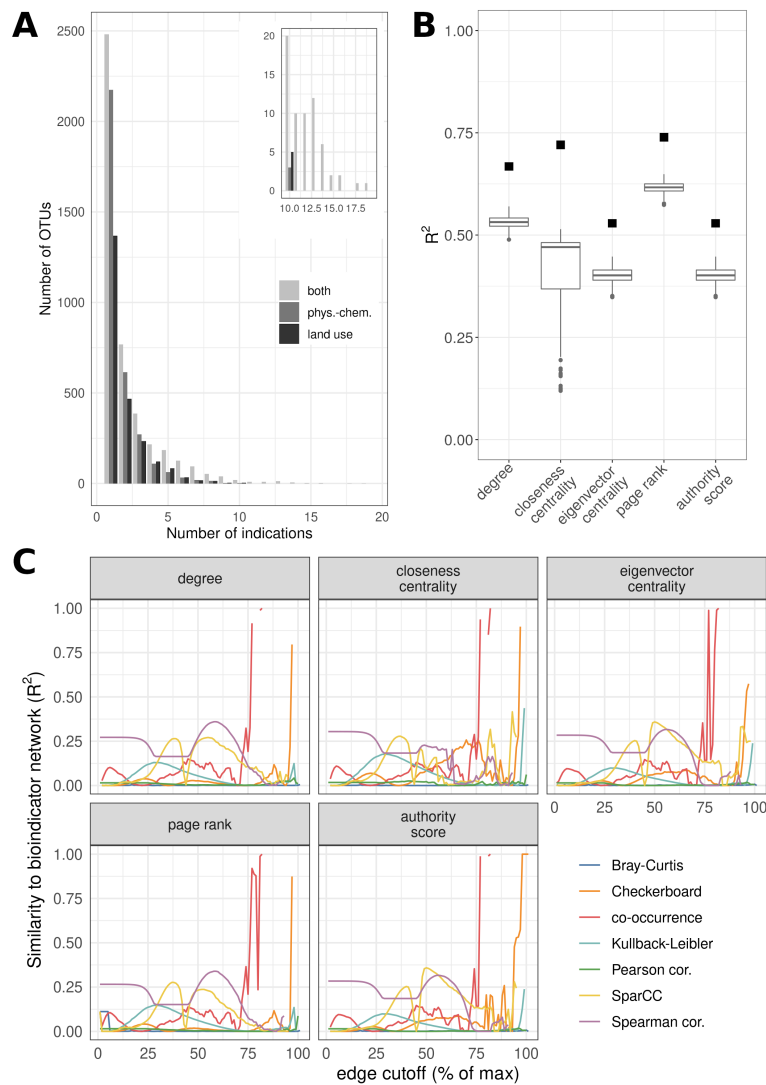


Figure 3: (Caption on the following page.)

Figure 3: (Previous page) The node properties of the bioindicator network are significant and comparable to co-occurrence networks. **A.** Distribution of multitask bioindicators across numbers of indicated land cover categories. Inset: Distribution of multitask bioindicators for ten or more parameters. **B.** Correlation between the cardinality of each bioindicator OTU and the node properties of the respective node in the bioindicator network. Black squares: Results for the bioindicator network. Grey box plots: Results for resampled null-hypothesis networks (for details, see methods). For all metrics, the results for the bioindicator network are significantly different from those of the null-hypothesis networks (one-sample t-test,  $P < 2.2e-16$ ). **C.** Comparison of the bioindicator network and established methods for the network inference from OTU tables. Networks are compared by the coefficient of determination,  $R^2$ , between a property of the nodes in the bioindicator network and the network created using a network inference method. As some of the methods create fully connected networks by default, edges with weights smaller than a cut-off were removed; this cut-off ranges between the respective minimum and maximum edge weight in 100 equidistant steps.

## Structural insights into the lake microbiome from multitask bioindicators

To further elucidate the relationship between the lake microbiome and environmental parameters, we combined the bioindicator OTUs identified for land cover parameters with those for physico-chemical parameters identified in (23). This way, we obtained a data structure that can be described as a set of maps between a set of OTUs and a set of environmental parameters. From this, we derived two distinct similarity matrices: One stating the similarity of OTUs in terms of the number of environmental parameters they are indicative of, and one that displays the similarity of the parameters in terms of the OTUs assigned to them.

The former can be turned into a bioindicator network as follows. Each bioindicator OTU is assigned to a node and edges are drawn between nodes representing OTUs that are indicative for at least one common environmental parameter (see supplementary table 3 for the entire network). We noticed correlations between the cardinality (i.e., the number of occurrences of an OTU in all bioindicator lists) of the nodes of the bioindicator network and the respective nodes' degree, closeness centrality, eigenvector centrality, page rank, and authority score (see supplementary figure 4). Because this result could be due to basic graph properties, we compared the square of the Pearson correlation coefficient,  $R^2$ , resulting from the correlation of cardinality with node properties of the bioindicator network with those gained from resampled networks (see Methods for details). We find that the nodes in the bioindicator network have statistically significant properties (see figure 3B), which suggests a biological relevance of the bioindicator network's structure.

Furthermore, we compared the bioindicator network to networks inferred from the original OTU table. More specifically, we asked whether the node properties generated using network inference methods correlate with the node properties of the bioindicator network. We chose this approach to comparing the two network structures as it can capture relative differences between node properties and might thus be robust with regard to global effects of a network method, e.g., a consistently lower degree. Our results show that applying a high cut-off to co-occurrence and checkerboard score similarity matrices results in networks similar to the

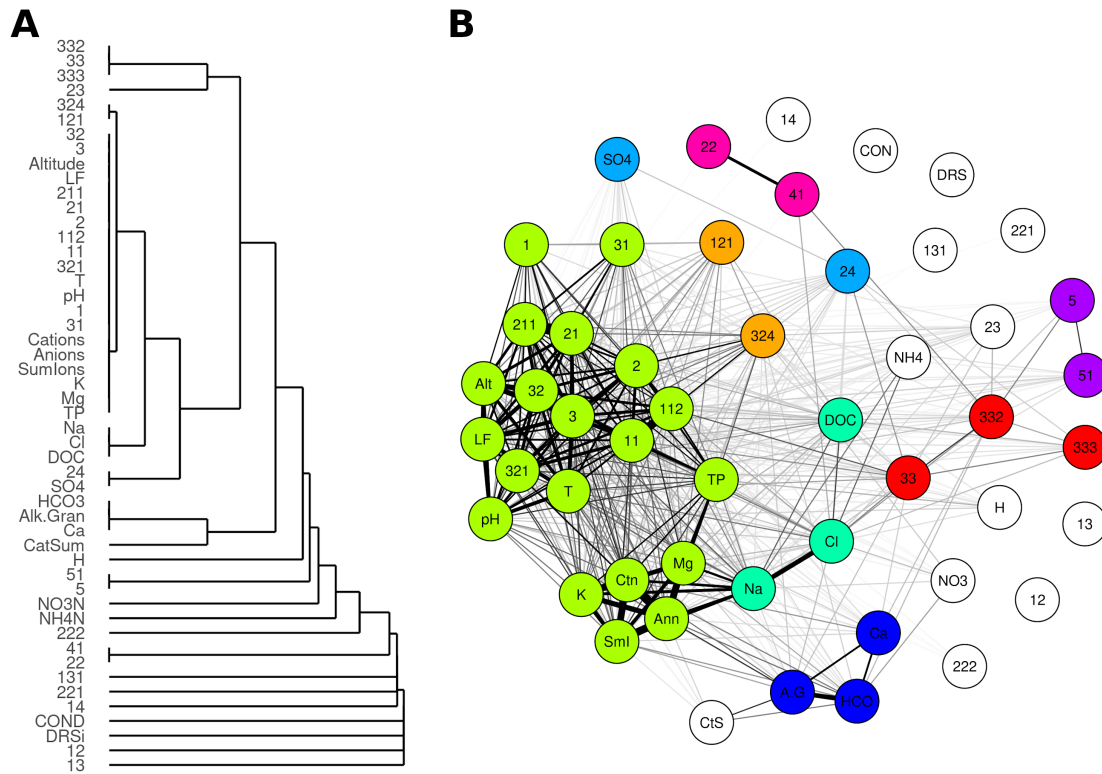


Figure 4: Clustering of physico-chemical and land cover parameters using the Jaccard similarity of the features' lists of bioindicator OTUs. Visualization of the similarity matrix as **A.** dendrogram using UPGMA and **B.** force embedded graph, in which edge size represents a higher Jaccard similarity. Node coloring in **B** represents clustering of parameters in **A.** Numbers represent CLC land cover categories according to the CLC legend (see table 1). Other abbreviations: A.G/Alk.Gran – alkalinity, Alt – altitude, Ann – anions, CatSum/CtS – sum of all cation concentrations, COND/CON – conductivity, Ctn – cations, DOC – dissolved organic carbon, DRSi/DRS – dissolved reactive silica, LF – conductivity (measured in the field, SumIons/SmI – sum of all ion concentrations, TP – total phosphorus.

bioindicator network. In contrast, neither correlation-based nor compositionality-aware methods do so (see figure 3C). These results suggest that organisms that strongly co-occur also tend to be indicative of the same environmental parameters.

The second data structure presents the similarity of pairs of environmental parameters in terms of the similarity of the bioindicator OTUs identified for them as given by the latter's Jaccard similarity. A high Jaccard similarity between two environmental parameters suggests that the microbiome responds to changes in the parameters in a similar manner. Along these lines, a visualization of this similarity matrix can be seen as a response map of the microbiome with regard to environmental changes. We attempted to visualize the resulting similarity matrix using non-metric dimensional scaling with up to 10 dimensions but were unable

to receive stress values  $< 0.05$ , suggesting that the responses of the microbiome to environmental change are non-trivial. Nevertheless, the visualization of the similarity matrix as a UPGMA-derived dendrogram and an undirected graph (see figures 4A and B, respectively) results in multiple distinct clusters of highly similar environmental parameters. The largest one of these comprises the concentration of magnesium, potassium, anions, cations, phosphorus, as well as temperature, pH, altitude, and a wide range of land cover categories. Furthermore, we observe a smaller cluster that contains the concentration of  $\text{HCO}_3$  and Calcium and the alkalinity; this might be due to the existence of a calcium-bicarbonate equilibrium in freshwater ecosystems (63) and a subpopulation of the lake microbiome responding to deviations from it. Additional information is needed to interpret the other clusters. A further notable result is the relatively high distances of the same CLC categories' subcategories in both the dendrogram and the graph. This underscores our prior finding that lake microbiomes react to different land cover categories in different ways (see figure 2).

## Discussion

The development of scalable Next-Generation Sequencing (NGS) methods has dramatically furthered the study of environmental microbiomes (64, 65, 66). However, technical and theoretical obstacles interfere with analyzing the wealth of data generated using NGS methods. For one, the dimensionality of microbiome datasets (i.e., the number of microbial species in ecosystems) is usually many orders of magnitude larger than the number of samples. For example, the dataset analyzed in this paper is, in terms of the number of samples, the biggest amplicon sequencing dataset for European lake microbiomes published to date, but still contains  $\sim 1000$  times more OTUs than samples. Together with the sparsity of OTU tables, this places the analyses of microbial communities at the edge of statistical feasibility, as, e.g., in such a domain, regression is ill-defined (67, 68, 69).

In a recent publication, we presented a machine learning-based approach to analyze the relationship between the microbiome and its environment (23). Training Random Forest models to approximate a projection of the microbial community composition to a one-dimensional space defined by one environmental parameter alleviates the issues that surround learning high-dimensional and sparse datasets containing non-linear feature dependencies (24). Furthermore, because the covariation framework employs the IndVal method (15) as a feature selection method before model training, the framework automatically identifies bioindicators for the environmental parameter in question.

In this paper, we use the covariation framework to study the relationship between lake microbiomes and the land cover surrounding the lakes. Our results underline the necessity for fine-grained analyses of the

impact of land use on freshwater ecosystems. While the predictability of microbial biodiversity based on land cover composition is consistently low over a range of radii (figures 1B and supplementary figure 3), we observe distinct covariations between separate land cover categories and the microbial lake community composition. This is true for all three levels of hierarchy in the CLC categorization of land cover (see figure 2 and supplementary table 1).

Our results suggest that the microbiome covaries to a considerable extent ( $R^2 > 0.3$ ) with the areas covered with forest-like vegetation (but not forest areas themselves), arable land, open spaces, plantations, and constructed environments such as urban areas and roads (see table 1, figure 2 and supplementary table 1). These results are in agreement with other recent studies of land cover and lake microbiomes (40, 70, 71). The low covariation of the microbiome with, e.g., areas of land covered with pastures or mine and construction sites (CLC categories 23 and 13, respectively) indicate that these types of land cover have no impact on the microbiome. Generally, the covariation between the microbiome and land cover categories are lower than those between the microbiome and physico-chemical parameters reported in ref. (23). This is the case because land cover does not impact lake microbiomes directly but is relayed via physico-chemical parameters. However, our results might be biased against variations in the land cover that are too small to appear in the CLC dataset but impact a lake’s ecology.

Just as the OTU counts, the environmental parameters studied here are not statistically independent but compositional and spatially autocorrelated (22). This leads to the emergence of parameters as drivers of ecological processes, i.e., parameters correlated with a high number of other environmental parameters without necessarily acting causally. In the context of this study, drivers appear as parameters with high numbers of indicators. Our analysis suggests the altitude of the lake is the primary driver of the microbial ecology of lakes (see table 1). The other parameters with high ( $> 500$ ) numbers of indicators have been described as drivers of lake ecology themselves (e.g., temperature) and/or known to be strongly correlated with altitude (as, e.g., herbaceous and forest vegetation, temperature, pollution, nutrient load), or with correlates of it (as water conductivity and pH is dependent on water temperature) (72, 73). Because the samples analyzed in this study were taken over a few days, we can exclude seasonal effects as confounders for our results.

In most settings, one would attempt to identify the effect of a single parameter on the study object by controlling for confounding effects via partial correlations or regression on residuals. However, spatial autocorrelation and indirect effects are significant for ecosystem functioning and integral for understanding it. They should not be considered “noise” that needs to be removed in analysis (19, 74, 75, 76). Instead, in

this study, we attempt to gain insight into how the microbiome responds to its environment by compiling the bioindicators for a range of environmental parameters. The resulting response map, visualized in figures 4B and C, is based on apparent (instead of direct, causal) relationships and depicts environmental parameters the microbiome responds to in a similar manner as connected with high-weighted edges. Such a response map holds great promise for the analysis of microbiomes as it can integrate heterogeneous environmental effects as long as the same microbiome is affected by them. The combined analysis of bioindicators identified for land cover and physico-chemical parameters performed here is proof for that.

Like environmental parameters that act as drivers of ecosystem processes, bioindicators indicative of a high number of environmental parameters, so-called multi-target bioindicators, might be considered central for the microbiome’s functioning. While experimental verification of the “keystone-ness” of multitask bioindicators is necessary (21, 77, 78, 79), the taxonomic distribution of the provisional multitask bioindicators for land cover categories identified here (see table 3) deserves a few words of discussion. First, the absence of Eukaryotes among the high-ranking multitask bioindicators suggests that bacterial niches can be more specific, making Bacteria more potent and more sensitive indicators for ecosystem health. Second, the relatively low number of Alpha- and Betaproteobacteria among the multitask indicators is in stark contrast to their high abundance in a broad range of freshwater ecosystems (80) as well as the environment-specific abundances of certain taxa among these classes (28). Both of these findings point towards the difference between fidelity and specificity as defined in the context of the IndVal method (13, 14, 15). Third and last, the attribution of ecological functionality to these OTUs is not possible, in part because only a tiny minority of microbes have been cultured and studied to a sufficient degree (81). In addition, we cannot assign correct species- or even strain-level taxonomic labels to OTUs of interest in this study because of the current limit of taxonomic resolution of amplicon sequencing and the incompleteness of taxonomic databases. Thus, how to attribute ecological function to microorganisms from amplicon datasets in a way that is comparable across studies and datasets remains one obstacle for studying environmental microbiomes.

## Acknowledgements

Calculations on the MaRC2 high-performance computer of the University of Marburg were conducted for this research. We would like to thank René Sitt of HPC-Hessen, funded by the State Ministry of Higher Education, Research and the Arts, for the installation and maintenance of software on the MaRC2 high-performance computer. We would like to thank Marius Welzel for helping with large-scale computing. This work was supported by the LOEWE program of the State of Hesse (Germany) in the MOSLA research

cluster. We also acknowledge funding by the Bauer-Foundation and Stemmler-Foundation for the project “Differential potential of metabarcoding, metatranscriptomics, and metagenomics for the assessment of lake water quality” and of the DFG project BO 3245/19-1.

## References

- [1] Allen CR, Angeler DG, Garmestani AS, Gunderson LH, Holling CS. Panarchy: Theory and Application. *Ecosystems*. 2014 1;17(4):578–589. Available from: <http://dx.doi.org/10.1007/s10021-013-9744-2>.
- [2] Dobzhansky T. Are Naturalists Old-Fashioned? *The American Naturalist*. 1966 9;100(915):541–550. Available from: <http://dx.doi.org/10.1086/282448>.
- [3] Odum EP. *Ecology - A Bridge Between Science And Society*. Sinauer Associates Incorporated; 1997.
- [4] O’Neill RV, Hunsaker CT, Jones KB, Riitters KH, Wickham JD, Schwartz PM, et al. Monitoring Environmental Quality at the Landscape Scale. *BioScience*. 1997 9;47(8):513–519. Available from: <http://dx.doi.org/10.2307/1313119>.
- [5] Williamson CE, Dodds W, Kratz TK, Palmer MA. Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. *Frontiers in Ecology and the Environment*. 2008 jun;6(5):247–254. Available from: <https://doi.org/10.1890/070140>.
- [6] Colwell RR. Microbial diversity: the importance of exploration and conservation. *Journal of Industrial Microbiology and Biotechnology*. 1997 5;18(5):302–307. Available from: <http://dx.doi.org/10.1038/sj.jim.2900390>.
- [7] Docherty KM, Gutknecht JLM. The role of environmental microorganisms in ecosystem responses to global change: current state of research and future outlooks. *Biogeochemistry*. 2011 sep;109(1-3):1–6. Available from: <https://doi.org/10.1007/s10533-011-9614-y>.
- [8] Webster NS, Wagner M, Negri AP. Microbial conservation in the Anthropocene. *Environmental Microbiology*. 2018 may;20(6):1925–1928. Available from: <https://doi.org/10.1111/1462-2920.14124>.
- [9] Birk S, Bonne W, Borja A, Brucet S, Courrat A, Poikane S, et al. Three hundred ways to assess Europe’s surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. *Ecological Indicators*. 2012 jul;18:31–41. Available from: <https://doi.org/10.1016/j.ecolind.2011.10.009>.
- [10] Hering D, Borja A, Carstensen J, Carvalho L, Elliott M, Feld CK, et al. The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. *Science of The Total Environment*. 2010 Sep;408(19):4007–4019. Available from: <https://doi.org/10.1016/j.scitotenv.2010.05.031>.
- [11] Cordier T, Forster D, Dufresne Y, Martins CIM, Stoeck T, Pawlowski J. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*. 2018 aug;18(6):1381–1391. Available from: <https://doi.org/10.1111/1755-0998.12926>.
- [12] Kermarrec L, Franc A, Rimet F, Chaumeil P, Frigerio JM, Humbert JF, et al. A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*. 2014 mar;33(1):349–363. Available from: <https://doi.org/10.1086/675079>.

- [13] De Cáceres M, Legendre P. Associations between species and groups of sites: indices and statistical inference; 2009. Available from: <http://sites.google.com/site/miqueldecaceres/>.
- [14] De Cáceres M, Legendre P, Moretti M. Improving indicator species analysis by combining groups of sites. *Oikos*. 2010 9;119(10):1674–1684. Available from: <http://dx.doi.org/10.1111/j.1600-0706.2010.18334.x>.
- [15] Dufrêne M, Legendre P. Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach. *Ecological Monographs*. 1997 aug;67(3):345–366.
- [16] Landres PB, Verner J, Thomas JW. Ecological Uses of Vertebrate Indicator Species: A Critique. *Conservation Biology*. 1988 12;2(4):316–328. Available from: <http://dx.doi.org/10.1111/j.1523-1739.1988.tb00195.x>.
- [17] Simberloff D. Flagships, umbrellas, and keystones: Is single-species management passé in the landscape era? *Biological Conservation*. 1998 3;83(3):247–257. Available from: [http://dx.doi.org/10.1016/S0006-3207\(97\)00081-5](http://dx.doi.org/10.1016/S0006-3207(97)00081-5).
- [18] Levin SA. Ecosystems and the Biosphere as Complex Adaptive Systems. *Ecosystems*. 1998 sep;1(5):431–436. Available from: <https://doi.org/10.1007/s100219900037>.
- [19] Levins R, Lewontin R. Dialectics and reductionism in ecology. *Synthese*. 1980 jan;43(1):47–78. Available from: <https://doi.org/10.1007/bf00413856>.
- [20] Green DG, Sadedin S. Interactions matter—complexity in landscapes and ecosystems. *Ecological Complexity*. 2005 jun;2(2):117–130. Available from: <https://doi.org/10.1016/j.ecocom.2004.11.006>.
- [21] Wang B, Zheng X, Zhang H, Xiao F, He Z, Yan Q. Keystone taxa of water microbiome respond to environmental quality and predict water contamination. *Environmental Research*. 2020 aug;187:109666. Available from: <https://doi.org/10.1016/j.envres.2020.109666>.
- [22] King RS, Baker ME, Whigham DF, Weller DE, Jordan TE, Kazyak PF, et al. Spatial Considerations for Linking Watershed Land Cover to Ecological Indicators in Streams. *Ecological Applications*. 2005;15. Available from: <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/04-0481>.
- [23] Sperlea T, Kreuder N, Beisser D, Hattab G, Boenigk J, Heider D. Quantification of the covariation of lake microbiomes and environmental variables using a machine learning-based framework. *Molecular Ecology*. 2021 3;0. Available from: <http://dx.doi.org/10.1111/mec.15872>.
- [24] Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5–32. Available from: <https://doi.org/10.1023/a:1010933404324>.
- [25] OpenStreetMap contributors. Planet dump retrieved from <https://planet.osm.org> ; 2017. <https://www.openstreetmap.org>.
- [26] European Union: Copernicus Land Monitoring Service. European Environment Agency (EEA). 2012;.
- [27] Boenigk J, Wodniok S, Bock C, Beisser D, Hempel C, Grossmann L, et al. Geographic distance and mountain ranges structure freshwater protist communities on a European scale. *Metabarcoding and Metagenomics*. 2018 jan;2:e21519. Available from: <https://doi.org/10.3897/mbmg.2.21519>.
- [28] Nuy JK, Hoetzing M, Hahn MW, Beisser D, Boenigk J. Ecological Differentiation in Two Major Freshwater Bacterial Taxa Along Environmental Gradients. *Frontiers in Microbiology*. 2020 feb;11. Available from: <https://doi.org/10.3389/fmicb.2020.00154>.
- [29] Bock C, Jensen M, Forster D, Marks S, Nuy J, Psenner R, et al. Factors shaping community patterns of protists and bacteria on a European scale. *Environmental Microbiology*. 2020 mar; Available from: <https://doi.org/10.1111/1462-2920.14992>.



- [30] Welzel M, Lange A, Heider D, Schwarz M, Freisleben B, Jensen M, et al. Natrix: a Snakemake-based workflow for processing, clustering, and taxonomically assigning amplicon sequencing reads. *BMC Bioinformatics*. 2020 nov;21(1). Available from: <https://doi.org/10.1186/s12859-020-03852-4>.
- [31] Andrews S. FASTQC. A quality control tool for high throughput sequence data; 2010.
- [32] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011 jan;27(6):863–864. Available from: <https://doi.org/10.1093/bioinformatics/btr026>.
- [33] Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*. 2012;13(1):31. Available from: <https://doi.org/10.1186/1471-2105-13-31>.
- [34] Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*. 2011 jun;27(16):2194–2200. Available from: <https://doi.org/10.1093/bioinformatics/btr381>.
- [35] Lange A, Jost S, Heider D, Bock C, Budeus B, Schilling E, et al. AmpliconDuo: A Split-Sample Filtering Protocol for High-Throughput Amplicon Sequencing of Microbial Communities. *PLOS ONE*. 2015 nov;10(11):e0141590. Available from: <https://doi.org/10.1371/journal.pone.0141590>.
- [36] Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*. 2014 sep;2:e593. Available from: <https://doi.org/10.7717/peerj.593>.
- [37] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology*. 1990 oct;215(3):403–410. Available from: [https://doi.org/10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2).
- [38] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*. 2012 Nov;41(D1):D590–D596. Available from: <https://doi.org/10.1093/nar/gks1219>.
- [39] QGIS Development Team. QGIS Geographic Information System; 2020. Available from: <https://www.qgis.org>.
- [40] Sperlea T, Füser S, Boenigk J, Heider D. SEDE-GPS: socio-economic data enrichment based on GPS information. *BMC Bioinformatics*. 2018 nov;19(S15). Available from: <https://doi.org/10.1186/s12859-018-2419-4>.
- [41] Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. *vegan: Community Ecology Package*; 2019. R package version 2.5-6. Available from: <https://CRAN.R-project.org/package=vegan>.
- [42] Chao A, Jost L. Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution*. 2015 feb;6(8):873–882. Available from: <https://doi.org/10.1111/2041-210x.12349>.
- [43] Daly A, Baetens J, Baets BD. Ecological Diversity: Measuring the Unmeasurable. *Mathematics*. 2018 jul;6(7):119. Available from: <https://doi.org/10.3390/math6070119>.
- [44] Kuhn M. Building Predictive Models in R using the caret Package. *Journal of Statistical Software*. 2008;28(5). Available from: <https://doi.org/10.18637/jss.v028.i05>.
- [45] Legendre P, Gallagher ED. Ecologically meaningful transformations for ordination of species data. *Oecologia*. 2001 oct;129(2):271–280. Available from: <https://doi.org/10.1007/s004420100716>.

- [46] Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*. 2012;48(4):1–18.
- [47] Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics*. 2011;27(4):592–593. Available from: <https://doi.org/10.1093/bioinformatics/btq706>.
- [48] de Vries A, Ripley BD. ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'; 2020. R package version 0.1.22. Available from: <https://CRAN.R-project.org/package=ggdendro>.
- [49] Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal*. 2006;Complex Systems:1695. Available from: <https://igraph.org>.
- [50] Connor EF, Simberloff D. The Assembly of Species Communities: Chance or Competition? *Ecology*. 1979;60(6):1132–1140. Available from: <http://www.jstor.org/stable/1936961>.
- [51] Stone L, Roberts A. The checkerboard score and species distributions. *Oecologia*. 1990 nov;85(1):74–79. Available from: <https://doi.org/10.1007/bf00317345>.
- [52] Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*. 1957 oct;27(4):325–349. Available from: <https://doi.org/10.2307/1942268>.
- [53] Kullback S, Leibler RA. On Information and Sufficiency. *The Annals of Mathematical Statistics*. 1951 mar;22(1):79–86. Available from: <https://doi.org/10.1214/aoms/1177729694>.
- [54] Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology*. 2012 sep;8(9):e1002687. Available from: <https://doi.org/10.1371/journal.pcbi.1002687>.
- [55] Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*. 2015 may;11(5):e1004226. Available from: <https://doi.org/10.1371/journal.pcbi.1004226>.
- [56] Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology*. 2014 may;5. Available from: <https://doi.org/10.3389/fmicb.2014.00219>.
- [57] Wickham H. ggplot2: Elegant Graphics for Data Analysis. vol. 0. Springer-Verlag New York; 2016. Available from: <https://ggplot2.tidyverse.org>.
- [58] Hattab G, Rhyne TM, Heider D. Ten simple rules to colorize biological data visualization. *PLOS Computational Biology*. 2020 10;16(10):e1008259. Available from: <http://dx.doi.org/10.1371/journal.pcbi.1008259>.
- [59] Gatti RC, Fath B, Hordijk W, Kauffman S, Ulanowicz R. Niche emergence as an autocatalytic process in the evolution of ecosystems. *Journal of Theoretical Biology*. 2018;454:110 – 117. Available from: <http://www.sciencedirect.com/science/article/pii/S0022519318302856>.
- [60] Martin G, Dang C, Morrissey E, Hubbart J, Kellner E, Kelly C, et al. Stream sediment bacterial communities exhibit temporally-consistent and distinct thresholds to land use change in a mixed-use watershed. *FEMS Microbiology Ecology*. 2020 12;97(2). Available from: <http://dx.doi.org/10.1093/femsec/fiaa256>.
- [61] Logares R, Tesson SVM, Canbäck B, Pontarp M, Hedlund K, Rengefors K. Contrasting prevalence of selection and drift in the community structuring of bacteria and microbial eukaryotes. *Environmental Microbiology*. 2018 may; Available from: <https://doi.org/10.1111/1462-2920.14265>.
- [62] Karlsson J, Jonsson A, Jansson M. Productivity of high-latitude lakes: climate effect inferred from altitude gradient. *Global Change Biology*. 2005 may;11(5):710–715. Available from: <https://doi.org/10.1111/j.1365-2486.2005.00945.x>.

- [63] Kopáček J, Hejzlar J, Oulehle F, Porcal P, Weyhenmeyer GA, Norton SA. Disruptions and re-establishment of the calcium-bicarbonate equilibrium in freshwaters. *Science of The Total Environment*. 2020 11;743:140626. Available from: <http://dx.doi.org/10.1016/j.scitotenv.2020.140626>.
- [64] Boughner LA, Singh P. Microbial Ecology: Where are we now? *Postdoc Journal*. 2016 nov;4(11). Available from: <https://doi.org/10.14304/surya.jpr.v4n11.2>.
- [65] Snyder LAS, Loman N, Pallen MJ, Penn CW. Next-Generation Sequencing—the Promise and Perils of Charting the Great Microbial Unknown. *Microbial Ecology*. 2008 nov;57(1):1–3. Available from: <https://doi.org/10.1007/s00248-008-9465-9>.
- [66] Tan B, Ng C, Nshimiyimana JP, Loh LL, Gin KYH, Thompson JR. Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges and future opportunities. *Frontiers in Microbiology*. 2015 sep;6. Available from: <https://doi.org/10.3389/fmicb.2015.01027>.
- [67] Carr A, Diener C, Baliga NS, Gibbons SM. Use and abuse of correlation analyses in microbial ecology. *The ISME Journal*. 2019 jun;13(11):2647–2655. Available from: <https://doi.org/10.1038/s41396-019-0459-z>.
- [68] Weiss SJ, Xu Z, Amir A, Peddada S, Bittinger K, Gonzalez A, et al. Effects of library size variance sparsity and compositionality on the analysis of microbiome data. *PeerJ PrePrints*. 2015;.
- [69] Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. 2017 mar;5(1). Available from: <https://doi.org/10.1186/s40168-017-0237-y>.
- [70] Kraemer SA, da Costa NB, Shapiro BJ, Fradette M, Huot Y, Walsh DA. A large-scale assessment of lakes reveals a pervasive signal of land use on bacterial communities. *The ISME Journal*. 2020 aug; Available from: <https://doi.org/10.1038/s41396-020-0733-0>.
- [71] Marmen S, Blank L, Al-Ashhab A, Malik A, Ganzert L, Lalar M, et al. The Role of Land Use Types and Water Chemical Properties in Structuring the Microbiomes of a Connected Lake System. *Frontiers in Microbiology*. 2020 feb;11. Available from: <https://doi.org/10.3389/fmicb.2020.00089>.
- [72] Forster D, Qu Z, Pitsch G, Bruni EP, Kammerlander B, Pröschold T, et al. Lake Ecosystem Robustness and Resilience Inferred from a Climate-Stressed Protistan Plankton Network. *Microorganisms*. 2021 3;9(3):549. Available from: <http://dx.doi.org/10.3390/microorganisms9030549>.
- [73] Urban D, Goslee S, Pierce K, Lookingbill T. Extending community ecology to landscapes. *Écoscience*. 2002 1;9(2):200–212. Available from: <http://dx.doi.org/10.1080/11956860.2002.11682706>.
- [74] Freckleton RP. On the misuse of residuals in ecology: regression of residuals vs. multiple regression. *Journal of Animal Ecology*. 2002 5;71(3):542–545. Available from: <http://dx.doi.org/10.1046/j.1365-2656.2002.00618.x>.
- [75] Legendre P. Spatial Autocorrelation: Trouble or New Paradigm? *Ecology*. 1993 9;74(6):1659–1673. Available from: <http://dx.doi.org/10.2307/1939924>.
- [76] Yodzis P. The Indeterminacy of Ecological Interactions as Perceived Through Perturbation Experiments. *Ecology*. 1988 4;69(2):508–515. Available from: <http://dx.doi.org/10.2307/1940449>.
- [77] Banerjee S, Schlaeppli K, van der Heijden MGA. Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology*. 2018 may; Available from: <https://doi.org/10.1038/s41579-018-0024-1>.
- [78] Banerjee S, Schlaeppli K, van der Heijden MGA. Reply to ‘Can we predict microbial keystones?’. *Nature Reviews Microbiology*. 2018 dec;17(3):194–194. Available from: <https://doi.org/10.1038/s41579-018-0133-x>.

- [79] Röttgers L, Faust K. Can we predict keystones? *Nature Reviews Microbiology*. 2018 dec;17(3):193–193. Available from: <https://doi.org/10.1038/s41579-018-0132-y>.
- [80] Šimek K, Kasalický V, Jezbera J, Horňák K, Nedoma J, Hahn MW, et al. Differential freshwater flagellate community response to bacterial food quality with a focus on Limnohabitans bacteria. *The ISME Journal*. 2013 apr;7(8):1519–1530. Available from: <https://doi.org/10.1038/ismej.2013.57>.
- [81] Thomas AM, Segata N. Multiple levels of the unknown in microbiome research. *BMC Biology*. 2019 jun;17. Available from: <https://doi.org/10.1186/s12915-019-0667-z>.

## Conflict of Interest

The authors declare that they have no competing financial interests.

## Data Accessibility

Raw sequencing data are available under the NCBI BioProject IDs PRJNA414052 and PRJNA559862.

## Author's contributions

TS designed and performed all computational analyses, JPS provided the OSM dataset, HD helped with the network analyses, DB performed the amplicon sequence analysis with Natrix, DB and JB provided the sequencing datasets, JB, GH, and DH supervised the study. All authors discussed the results and wrote and revised the manuscript.

## Tables

Table 1: Numbers of microbial indicators for land cover categories (with respective radius) and physico-chemical parameters and the respective  $R^2$  resulting from the covariation framework (results for physico-chemical parameters taken from (23)).

Parameter	Indicators	Radius (km)	$R^2$	Parameter	Indicators	$R^2$
Herbaceous vegetation (32)	1056	7	0.42	Altitude	1595	0.60
Forest, semi-natural areas (3)	703	10	0.36	Conductivity (LF)	1349	0.49
Arable land (21)	445	9	0.36	Temperature	920	0.54
Non-irrigated arable land (211)	416	9	0.34	pH	603	0.31
Discontinuous urban fabric (112)	276	8	0.29	Sum of Ions	166	0.50
Agricultural areas (2)	272	-	-	Cations	164	0.52
Natural grasslands (321)	265	10	0.29	K <sup>+</sup>	118	0.26
Urban fabric (11)	247	7	0.29	Total Phosphorus (TP)	92	0.31
Artificial surfaces (1)	177	-	-	Anions	89	0.53
Forests (31)	140	7	0.15	Na <sup>2</sup>	86	0.18
Inland waters (51)	77	6	0.18	Mg <sup>2</sup>	70	0.34
Industrial or commercial units (121)	77	10	0.17	DOC	43	0.40
Heterogeneous agricultural areas (24)	69	-	-	Ca <sup>2</sup>	32	0.37
Transitional woodland-shrub (324)	61	10	0.19	HCO <sub>3</sub>	27	0.40
Open spaces with little vegetation (33)	52	9	0.42	NO <sub>3</sub>	17	0.29
Bare Rocks (332)	34	9	0.32	Alkalinity (Alk.Gran)	16	0.38
Water bodies (5)	32	-	-			
Pastures (23)	21	-	-			
Sparsely vegetated areas (333)	15	2	0.38			
Permanent crops (22)	6	-	-			
Fruit trees (222)	3	8	0.40			
Inland wetlands (41)	2	4	0.20			
Artificial, non-agricultural vegetated areas (14)	2	-	-			
Industrial, commercial and transport units (12)	2	-	-			
Vineyards (221)	1	6	0.18			
Mineral extraction sites (131)	1	9	0.18			
Mine, dump and construction sites (13)	1	9	0.18			
Dump sites (132)	1	8	0.22			
Construction sites (133)	1	10	0.57			

Table 2: Multitask OTUs identified in this study for more than 6 land cover categories and the physico-chemical parameters they have been identified as indicators for in (23). Numbers represent CLC land cover categories (see table 1). For taxonomic annotation of these OTUs, see table 3.

	Freq	X1	X11	X112	X121	X2	X21	X211	X23	X24	X3	X31	X32	X321	X324	X33	X332	Phys.-chem.
N1115	10		X	X	X	X	X	X	X		X		X	X	X			Altitude LF T
N213	10		X	X	X	X	X	X			X		X	X	X			Altitude LF Mg pH T
N35	10	X	X	X	X	X	X	X			X		X	X	X			Altitude LF pH T
N427	10		X	X	X	X	X	X			X		X	X	X			Altitude K LF Mg
N74	10	X	X	X	X	X	X	X			X		X	X	X			Altitude LF TP
N1077	9	X	X	X	X	X	X	X			X		X	X	X			Altitude Cations K LF Mg Na pH SO4 SumIons T
N178	9	X	X	X	X	X	X	X			X		X	X	X			Altitude LF pH T TP
N305	9	X	X	X	X	X	X	X			X		X	X	X			Altitude LF
N469	9	X	X	X	X	X	X	X			X		X	X	X			Altitude LF
N1276	8		X	X	X	X	X	X			X		X	X	X			Altitude LF
N156	8					X	X	X			X		X	X	X			Altitude CI DOC T
N166	8		X	X	X	X	X	X			X		X	X	X			Altitude LF pH T TP
N177	8		X	X	X	X	X	X			X		X	X	X			Alk.Gran LF
N1844	8		X	X	X	X	X	X			X		X	X	X			Altitude Anions Cations K LF pH T TP
N243	8					X	X	X			X		X	X	X			Altitude LF T
N470	8	X	X	X	X	X	X	X			X		X	X	X			Altitude K LF pH T TP
N471	8					X	X	X			X		X	X	X			Altitude LF pH
N513	8		X	X	X	X	X	X			X		X	X	X			Altitude LF pH T TP
N519	8		X	X	X	X	X	X			X		X	X	X			Altitude Anions Cations K LF Mg pH SumIons T TP
N533	8		X	X	X	X	X	X			X		X	X	X			Altitude LF T
N563	8		X	X	X	X	X	X			X		X	X	X			Altitude Anions Cations LF pH SumIons T TP
N689	8		X	X	X	X	X	X			X		X	X	X			Altitude K LF pH T TP
N785	8		X	X	X	X	X	X			X		X	X	X			Altitude LF pH T TP
N98	8		X	X	X	X	X	X			X		X	X	X			Altitude LF pH T

Table 3: Taxonomic annotation of the multitask OTUs presented in table 2.

	Domain	Phylum	Class	Order	Family	Genus	Species
N1115	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	<i>Sphingomonas</i>	<i>Sphingomonas</i> sp. CGGE4131
N1213	Bacteroidetes	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cycomorphaceae	<i>Cycomorphus</i>	
N35	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	<i>Roseomonas</i>	clone B197(2011)
N427	Bacteroidetes	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cycomorphaceae	<i>Flavobacterium</i>	
N74	Bacteria	Bacteroidetes	Cytophagia	Cytophagales	Cycomorphaceae		
N1077	Bacteroidetes	Bacteroidetes	Cytophagia	Cytophagales	Cycomorphaceae		
N178	Bacteroidetes	Bacteroidetes	Flavobacteriia	Flavobacteriales	NS9 marine group		
N305	Bacteroidetes	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cycomorphaceae	<i>Flavobacterium</i>	
N469	Bacteroidetes	Bacteroidetes	Betaproteobacteria	Burkholderiales	Comamonadaceae		
N1276	Bacteria	Proteobacteria	Cyanobacteria	Subsection I	Family I		
N156	Bacteroidetes	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cycomorphaceae	<i>Synechococcus</i>	
N166	Bacteroidetes	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cycomorphaceae	<i>Flavobacterium</i>	
N177	Bacteroidetes	Bacteroidetes	Flavobacteriia	Flavobacteriales	NS9 marine group		
N1844	Firmicutes	Firmicutes	Clostridia	Clostridiales	Clostridiaceae 1	<i>Clostridium</i> sensu stricto 1	
N243	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacteriales	Rhodobacteraceae	<i>Rhodobacter</i>	
N470	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	<i>Acidibacter</i>	
N471	Bacteroidetes	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	<i>Flavobacterium</i>	
N513	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	<i>Acidibacter</i>	<i>Flavobacterium</i> sp.
N519	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cycomorphaceae	<i>Flavobacterium</i>	
N533	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	<i>Novosphingobium</i>	
N563	Bacteroidetes	Bacteroidetes	Opitutae	Opitutales	Opitutaceae	<i>Opitutus</i>	
N689	Bacteria	Verrucomicrobia	Sphingobacteriia	Sphingobacteriales	NS11-12 marine group	uncultured Bacteroidetes bacterium	
N785	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	NS9 marine group	uncultured soil bacterium	
N98	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	<i>Flavobacterium</i>	

#### 2.4 PUBLICATIONS NOT INCLUDED IN THIS THESIS

Hannah F. Löchel, Dominic Eger, Theodor Sperlea, and Dominik Heider (2019). Deep learning on chaos game representation for proteins. *Bioinformatics*, 36(1):272–279

Theodor Sperlea, Lea Muth, Roman Martin, Christoph Weigel, Torsten Waldminghaus, and Dominik Heider (2020). gammaBOriS: Identification and taxonomic classification of origins of replication in Gammaproteobacteria using motif-based machine learning. *Scientific Reports*, 10(1):6727



# 3

## Discussion

### 3.1 INTRODUCTORY REMARKS

The work presented in this thesis centers around the obstacles one encounters when studying complex systems in an observational instead of an experimental setting. More specifically, the focus of this thesis is the lake microbiome, which is a relatively new study object. This chapter will refrain from repeating points of discussion already covered in the publications included herein. Instead, I will write on two questions that imposed themselves on me when I was putting together this thesis – and survived the scrutiny of being relevant to the results presented herein. This chapter is structured as follows. First, I will ask what the necessary conditions for the assumption of comparability to be true are and whether I can, *ex post*, whether the lakes studied in the publications are, in fact, sufficiently similar. After that, I will show that what is estimated by the covariation framework is a measure of coupling in complex systems by proposing a theory of information for biomonitoring. To do so adequately, however, requires an information theory of ecosystems, which is why section 3.3 is the longest in this chapter. In section 3.4 I will then re-motivate the covariation framework and discuss potential methodological improvements. Finally, section 3.5, will discuss how to build on these results – and what might be relevant for the study of environmental microbiomes in the future.

### 3.2 THE COMPARABILITY OF MICROBIOMES

Out of the assumptions made explicit in section 1.4.2, assumptions I to III stand out as worthy of more discussion. All three are, in essence, assumptions of comparability; comparability of samples, ecosystems, and microbiomes, respectively. Statistical analyses and machine learning methods require a certain degree of similarity in the data, even if used to determine differences. These approaches presuppose the idea of meaningfully defined populations of data points, for which we can calculate averages and standard deviations. The following will not discuss how batch effects or changes in the sequencing analysis pipeline affect the comparability of OTU tables<sup>110</sup>. Instead, I want to focus on how the complexity of the objects studied here affects comparability.

For objects to be comparable, they need to have a large enough set of properties in common. Otherwise, any comparison will result in nonsensical statements. For example, we can compare apples and oranges and find meaningful similarities as well as characteristic distinctions: They are both fruits but exhibit differences in color, shape, and taste. In contrast, the same is almost impossible when comparing apples to, say, directives in medieval legislation because these two sets of objects lack even the smallest amount of common ground. It is surprisingly hard to formulate exact conditions of comparison. Suffice it to say, objects may be compared with merit if and only if they already resemble each other to a certain degree or share a certain amount of properties<sup>52</sup>.

But what are the limits of comparability when it comes to environmental microbiomes? Because different ecosystem types, such as freshwater, soil, or forest ecosystems, harbor distinct microbiomes<sup>115,120</sup>, comparisons across ecosystem types will, most likely, not be helpful. For example, the microbiomes present in sediments or the water of streams are highly distinct<sup>332</sup>. In soil ecosystems, comparability is further complicated by a high degree of spatial diversity, partly due to land cover and land use, partly due to soil patchiness<sup>176,245</sup>. In contrast, we expect a more (albeit not completely) homogenous microbiome throughout different points in a lake because of water circulation<sup>120</sup>. Along the same lines, many studies have shown that the dynamics of microbiomes follow seasonal patterns<sup>93,95,102,111,147,154,176,186</sup>, which can complicate the comparison of microbiome samples taken from the same sampling site in different seasons. Of course, this depends on the research question at hand: If one were to study the seasonality of microbiome dynamics, long-term sampling is indispensable.

Complexity, as described in section 1.2.1, introduces further barriers for comparability in that ecosystems can be in different regimes or attractors; on different sides of a regime shift or tipping point, lakes might not resemble one another to a sufficient degree<sup>118</sup>. For example, the dynamics in a lake will change drastically after eutrophication, to the degree that we might conclude that it is a “completely different beast”<sup>22,69</sup>. If the European lake dataset contained samples from lakes that belong to two separate attractors, we would, in statistical terms, have two different populations in our dataset with distinct dynamics. We would need to abstain from doing statistical analyses with them; doing so would be analogous to attempting to find a relationship between the amount of watering a plant has enjoyed throughout its lifetime and the size of its fruit, but not distinguishing between apple and orange trees.

The design of the sampling schemes and initial analyses did not suggest that there would be multiple distinct populations of lakes in different regimes in the datasets. However, there is a slight allusion in my results that the lakes in the European dataset are not strictly comparable: While we find almost perfect predictability of microbial biodiversity from socio-economic and land cover data in publication I (figure 2, section 2.1), the predictability of microbial biodiversity from land cover data is low in publication III (figure 1, section 2.3). This discrepancy is not discussed in any of the papers because the two cases of machine learning are, strictly speaking, not comparable in their own right: In the first, we used a very high number of features, i.e., the entire output of SEDE-GPS, whereas, in the second, only OpenStreetMap-derived land cover data was available to the model. Thus, the lower number of features available for machine learning in publication III might be the reason for the poor performance of the models in this case.

However, the stark contrast between these two calculations might also be due to differences in the dataset. Publication I investigates a set of alpine lakes with high geographical proximity. In contrast, the results in publication III are based upon the whole European lake dataset, combining lakes separated by large distances. The difference might stem from the fact that the hypothesis “Everything is everywhere” does not hold, i.e., the geographic dispersal of microbes is not uniform but partially shaped by geographic barriers<sup>30,120,231</sup>. While, after all, I maintain that comparability is warranted in the European lake dataset by way of study design and outlier detection and removal, deviations of some lakes from the common dynamics of the others might lead to lower  $R^2$  values in analyses like these.

It is more worrisome that my results are not directly comparable with those derived from other sampling sites or other datasets. There are two main reasons for this. For one, the OTUs identified in the raw metabarcoding data by the tool SWARM<sup>30,196,317</sup> are generated by clustering of the sequences present in the dataset and, thus, not independent of its structure. As such, we cannot compare OTU tables generated separately from different datasets. One might avoid this issue when creating ASVs instead of OTUs (for details, see section 1.2.4)<sup>40</sup>. However, this does not solve the second barrier, i.e., the incompleteness of taxonomic reference databases, which leads to an inability to taxonomically annotate a large proportion of OTUs<sup>265</sup>. Both the definition of OTUs and the incompleteness of references lead, furthermore, to a below-par taxonomic resolution – that is to say, while it is usually possible to assign an OTU to a Class or Order, this is not the case below the Family level. Incomplete taxonomic annotation leads to rather non-descript lists of OTUs, such as present in table 3 of publication II (section 2.2), table 3 in publication III (section 2.3) as well as in publications from other groups<sup>5,101</sup>. Making the results of studies of environmental microbiomes more comparable by removing technical barriers while locating natural barriers of comparability will significantly fuel this field’s progress.

### 3.3 A THEORY FOR MICROBIAL BIOMONITORING

A central point of this thesis is the covariation between the microbiome as a whole and separate environmental parameters. However, until this point, covariation has remained an abstract measure without a clear interpretation. In the following, I will develop a theory of biomonitoring schemes that make use of complex assemblages like microbiomes and make clear that the covariation is a measure of coupling in complex environments. In short, my argument will go as follows: In biomonitoring, the microbiome can be described as a biosensor, i.e., somewhat similar to technical measuring equipment, but with some significant characteristics that derive from complexity (section 3.3.1). The classical information theory put forward by Claude Shannon cannot adequately describe the way that the microbiome captures information about the ecosystem (3.3.2). An alternative, extended theory of information that centers around a notion of pragmatic information can do so (section 3.3.3) but requires us to think about the environment of the microbiome in a very definite and strong sense (section 3.3.4). From that vantage point, it becomes clear that the covariation framework estimates a measure of coupling between disjoint systems (or assemblages). To make this argument convincingly, I combine approaches from different

disciplines which come with their dedicated nomenclatures. While I tried to use consistent terminology, this was not possible at all steps of the argumentation; I hope the reader will appreciate the irreducible diversity of positions involved in the section.

### 3.3.1 THE MICROBIOME AS A BIOSENSOR

As suggested in section 1.3.3, the canaries used in coal mines to warn the miners of lethal gases represent an archaic form of biomonitoring: Here, we have a living system acting as a sensor for a parameter that is not easily measurable in its own right. At that, the canary offered a clear read-out with a binary value: As long as the canary is alive, the air is clear; else, one should leave the mine in danger of meeting the same fate as the canary. In analogy, in ecological biomonitoring, observables (i.e., sufficiently easily measurable properties) of the monitored organism(s) are used as proxy measurements for environmental parameters such as physico-chemical parameters or ecosystem health (as described in section 1.3.3). Note that it does not, for now, make a difference whether the sensor used for biomonitoring (in the following called bio-sensor) is a single species of, say, lichens, a larger set of morphologically determined diatoms, or a whole microbiome. It is sufficient for the organism(s) to be at a Pareto optimal point regarding the accuracy of the approximation of the target variable on the one hand and the ease of extracting environmental information from them on the other.

Let us, furthermore, clarify the contrast between biosensors and “anthropogenic sensors”, measuring devices such as thermometers, kitchen scales, and Geiger counters. Biosensors are, necessarily, *objets trouvés*, i.e., ready-made, pre-existing, and discovered living systems or assemblages of living systems. Unlike the well-constructed measuring devices, whose whole purpose is to be a sensor, the internals “connectome” of a biosensor is, usually, not well-understood. Because of this, we need to handle a biosensor as a black box with the parameter it will be a proxy for as input and the observable(s) it presents as an output. Furthermore, to be able to use organisms as functional sensors, we need something akin to a user interface – usually, a mathematical model representing the observables’ values as something understandable. In essence, this role is played by biotic indices and, more recently, machine learning models.

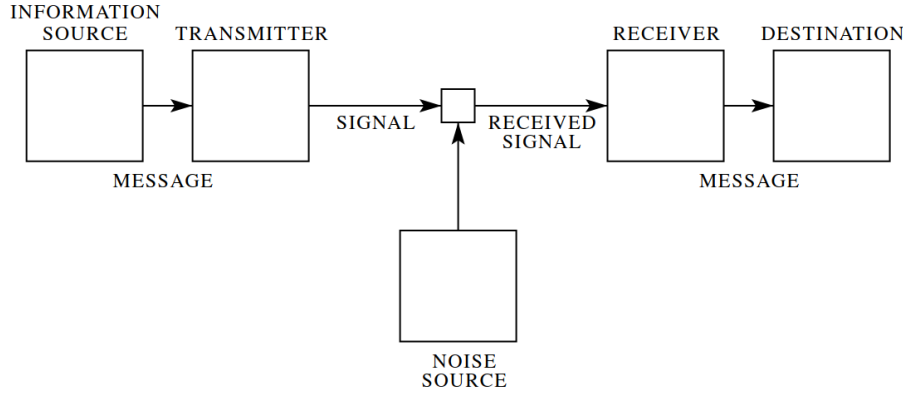
It is important to stress that, for biomonitoring alone, we can be ignorant about the way the sensor works as long as it does do so. That is to say, the relationship between these two values does not need to be functional, causal, or direct, as long it is apparent and reliably reproducible.

This limitation is explicit in, for example, the IndVal function (see section 1.3.3), whose authors insist that their method for identifying bioindicators is no determinant of causal connections<sup>60</sup>. Furthermore, because of the high degree of interconnection in complex environments, it is not important which kind of observable we use – may it be the population size, the microbial community composition, or the metatranscriptome in the case of microbial biomonitoring – as long as it has an apparent relationship to the parameter in question. If we were also interested in ecosystem management, this would not suffice, as causal models would be necessary to guide interventions into the ecosystem. However, the identification of causal relationships in complex systems is a question far from being answered except in particular situations<sup>140,291,300,329</sup>. Therefore, for the sake of this argument, we will stick to biosensors for biomonitoring and their anthropogenic counterparts and ask how the sensor has access to the parameters in question.

### 3.3.2 INFORMATION THEORY

One could describe how a biosensor functions as a process of information transfer: The environmental parameters contain information on the current state of the ecosystem – and by sensing these parameters, the sensor can reflect this information in its observables. However, the rather general and potentially opaque meaning of information<sup>142</sup> might confuse and obfuscate essential details of how biosensors work. Thus, it is necessary to review and adopt a definition of information in line with biomonitoring to arrive at a coherent theory of microbial biosensors.

The most widely used of these can be traced back to a seminal paper by Claude E. Shannon published in 1948<sup>279</sup>. Its groundlaying model can be described as follows (see figure 3.1): A message is sent by an information source and encoded by a transmitter before the resulting signal is passed through a potentially noisy channel. The receiver decodes the signal so that the (original) message can arrive at its destination. Often, in the literature building on Shannon's original paper, the information source is called "Alice", while the destination is named "Bob". It is important to stress that Alice and Bob are not (necessarily) conscious, human communicators. In fact, Shannon's "mathematical theory of communication" is not a theory of human communication but presents a set of mathematical formulae that enable the construction of transmitters, channels, and receivers in such a way as to transmit information faithfully in the presence of noise. Furthermore, this theory has proven to be greatly useful for electronic communication between computers, cryptography, and information storage systems, all of which do not necessarily in-



**Figure 3.1:** The model of information transfer in the “Mathematical Theory of Information” by Claude Shannon, taken from Shannon (1948)<sup>279</sup>. For details, see main text.

volve humans as sources or destinations of information. I will adopt this convention throughout this and the following section to reduce the technical jargon to the necessary amount.

For communication as modeled by Shannon to take place, Alice and Bob need to agree, in advance, on what kind(s) of messages will be transmitted. For example, they might want to restrict the set of possible messages to “a common vocabulary”<sup>170</sup>, as, e.g., the letters in the Latin alphabet, ones and zeroes in the case of binary messages, or continuous values in a specified range. The same goes for the type of signal sent through the channel and the encoding function implemented in the transmitter so that the receiver can implement the inverse of it and decode signals received, after additional error correction steps, into “meaningful” messages. These presuppositions are usually implicit and obvious for electronic communication systems but crucial if we apply this notion of information to a different field.

Given this set-up, Shannon proposes a formalism with a set of information measures, of which only I will discuss a small subset here. The most central of these is the amount of information present in a sequence of  $n$  independent symbols, given in analogy to the thermodynamic entropy by

$$H_{discrete} = - \sum_i^n p(i) \log_2(p(i)), \quad (3.1)$$

where  $p(i)$  represents the probability of the  $i$ th symbol of the message. As mentioned before, the

alphabet  $\mathbb{X}$  must be predefined so that

$$\sum_{x \in \mathbb{X}} p(x) = 1. \quad (3.2)$$

In other words, Shannon models the information source like a stochastic process and is most interested in the deviations from the probability distribution  $p_x$  or  $p(x)$ . It compares the sequence of information actually sent to all other possible sequences to arrive at a measure of information. As such, the Shannon information is a metric of syntactic surprisal: Before receiving the next symbol, Bob's expectations (or, in analogy to Bayesian statistics, prior) of the upcoming symbol are equal to  $p_x$ , and the occurrence of rare symbols leads to a larger increase in  $H_{discrete}$  than the occurrence of symbols with large  $p_x$ .

Especially when facing a noisy channel, a somewhat more important measure of information proposed by Shannon is the mutual information (or transinformation) of  $X$ , the message as sent by Alice, and  $Y$ , the message as received by Bob after going through the channel. More formally, it quantifies the degree of dependence between  $X$  and  $Y$  and can be given by

$$I(X; Y) = D_{KL}(p(x, y) \parallel p(x)p(y)) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}, \quad (3.3)$$

where  $p(x, y)$  is the joint probability of  $X$  and  $Y$ , while  $p(x)$  and  $p(y)$  represent marginal probability functions of  $X$  and  $Y$ , respectively, and  $D_{KL}$  denotes the Kullback-Leibler divergence. Because if  $X = Y$ , then  $I(X; Y) = 1$ , this measure also gives us an indication to the amount of noise affecting the channel provided we can access both  $X$  and  $Y$  at the same time.

Because Shannon was able to show mathematically that reliable communication is possible via unreliable or noisy channels given a high enough channel capacity, his “mathematical theory of communication” and its formalism laid the groundwork for later development in electronic communication systems. Furthermore, the intuitiveness of its basic model (see figure 3.1) and its terminology led to a quick adoption in other fields of study – and even a considerable level of popularity in those fields in which its formalism is not applicable. The somewhat quick and unguarded adaptation of Shannon's theory by other disciplines has been criticized by many authors<sup>83,142,189,312</sup>, including Shannon himself, who, in a opinion piece from 1956 described this as a

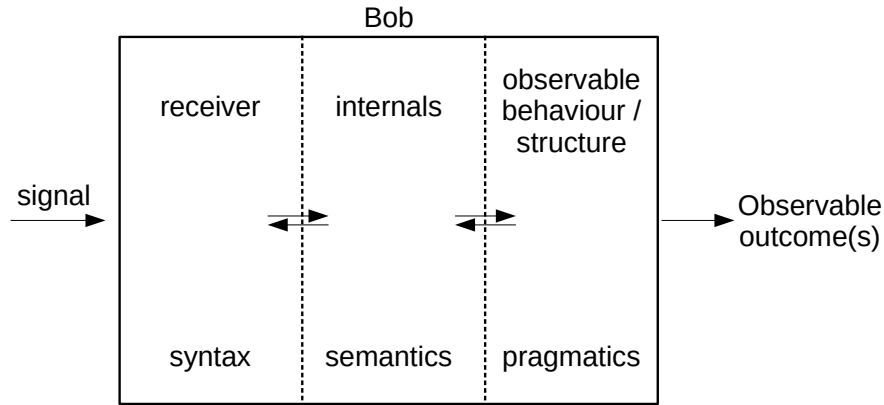


“bandwagon [that] has perhaps been ballooned to an importance beyond its actual accomplishments”, warning that “a few exciting words like information, entropy, redundancy, do not solve all our problems” and admitting that “[t]he subject of information theory has certainly been sold, if not oversold.”<sup>280</sup>

Ironically, those aspects of Shannon’s theory that make it well-suited for its original use case are also those aspects that limit its usefulness for biomonitoring. For one, this information theory is closely linked to the model given in figure 3.1: All the parts depicted there need to be present or identifiable in the situations we would like to apply it. This model implicitly entails a higher degree of autonomy or activity for Alice than for Bob: A message needs to be (actively) sent by the information source and cannot be fetched by the message’s destination. In the original publication, Shannon spends more time on specifying the information source than on the destination, which he laconically describes as “the person (or thing) for whom the message is intended”<sup>279</sup> as if taking its existence for granted. However, given a sensor, it would be strange to say that the temperature in the room, a kilo of flour, or the decaying radioactive material are somewhat actively sending a message to the measuring device we use as a sensor. The same goes for the microbiome as biosensor: Only with considerable contortions one can identify an environmental parameter or the ecosystem itself as the sender, just as there is no message, de- or encoder, or even channel in a sense that would be a good representation of real ecosystems and also fit the model of the “mathematical theory of communication”. Finally, Shannon explicitly excludes questions of meaning as a reasonable subject of an information theory by stating that the “semantic aspects of communication are irrelevant to the engineering problem”<sup>279</sup>, but, as will be shown below, the semantic aspect is crucial for a theory of information for biomonitoring.

### 3.3.3 AN EXTENDED VIEW OF INFORMATION

As we have seen, Shannon information has some *a priori* limitations and assumptions that are necessary for its use case but problematic when applied to other disciplines. An even worse fit is Kolmogorov’s theory of algorithmic information, which would denote the shortest program to produce a given object<sup>165,189</sup>. Lesser known, more encompassing, and, as I will argue, more realistic when it comes to sensors and biomonitoring is the tripartite notion of information derived from semiotics. In spite of their history, I will avoid analogies to natural languages as I am not sure whether these analogies hold. In general terms, the tripartite notion of information can be



**Figure 3.2:** A graphical model for the tripartite notion of information as applied to communication. For details, see main text.

outlined as follows<sup>7,54,55,208</sup>:

1. Syntactic information is the information present in the order of symbols in the message.
2. Semantic information is the meaning of the message for Bob.
3. Pragmatic information is the effect the message has on the state, behavior, or choice of actions of Bob.

In accordance with this enumeration, we can divide Bob's structure into three interdependent and interlinked parts: the receiver, the latent structure, and the observable structure (figure 3.2). Furthermore, all three of these aspects are dependent on the current state of Bob, and, as will be clear from what follows, the pragmatic information is the best possibility to measure information in complex systems<sup>170</sup> – in fact, the other aspects are not measurable at all<sup>17</sup>.

It is a consensus shared by many theorists of information that a system only carries information if it can be in at least two states and it is in one but not the other<sup>20,56,57,70,195,320</sup>. This reduces to Shannon information when applied to a pre-defined set of possible states: Each position of a string might be filled by any one of the characters in an alphabet but is only occupied by one of them. Thus, Shannon information is a special case of syntactical information. However, we need

to adopt a broader one for use in the context of complex systems. In However, the same is not the case when the system that carries the information is complex, as there is usually more than one way of distinguishing between states of a complex system<sup>55</sup>. Has a complex system shifted into another state if one of its parameters has changed gradually or is a larger-scale change necessary? This distinction is not trivial and might not be universal. If we now take Alice to be such a complex system, this insight shifts the task of defining what might act as syntactical information to Bob. More specifically, Bob needs to distinguish between changes in the system that are minute (and, therefore, ignorable noise) and changes that represent information to Bob (and, thus, a signal to receive)<sup>57,80,142</sup>. While I make use of anthropomorphic language, it is important to reiterate that Bob is not necessarily a human or a conscious, living organism. Kitchen scales, for example, distinguish between relevant and irrelevant information by way of their measuring accuracy: Changes in weight too small are simply not registered by its electronic internals. Thus, the selection of syntactic information is dependent on the structure of Bob<sup>100,193,195</sup>.

After receiving the signal, Bob translates the syntactical information into semantic information, which, for brevity and simplicity and as I did above, is often paraphrased as meaning. Of course, doing so is potentially misleading, especially in the case of a Bob that does not implement cognitive processes. Instead, we should picture semantic information as the internal changes (changes to the “state of conditional readiness”<sup>195</sup>) in Bob that occur when the signal is received. Going back to the kitchen scale, we can identify the semantic aspect of information with the electronic pulses induced in the sensor. Similarly, we can equate the meaning of the binding of a small molecule to a receptor on a bacterial cell wall to the activation of some specific pathways while others are not activated and the subsequent differential changes in gene expression inside the cell. The semantic aspect of information is strongly linked to the internal structure of Bob – so much so that in those cases where we do not fully understand the internal wiring of Bob, we are unable to access this aspect of information and, therefore, unable to quantify it.

The pragmatic aspect of information can be defined as the observable effect of the signal received by Bob<sup>166,170,312,311</sup>. Because the others are not measurable from the outside (except for in particular cases), this might be the only aspect of information relevant for the scientific endeavor<sup>6,308,167,312</sup>. That is to say, in the words of Gregory Bateson, “the elementary unit of information – is a difference that makes a difference”<sup>21</sup>. The intuition behind this is that if a signal does not lead to a change of behavior or structure in Bob, it is either irrelevant (i.e., Bob does not act

as receiver of the signal) or synonymous with another signal Bob is receiving<sup>167</sup>. In the case of the microbiome as a biosensor, we might interpret changes in its community structure as an indication of pragmatic information. Furthermore, given the appropriate amount of care, it will not matter whether we use OTU tables, metatranscriptomes or any other *meta-omics* dataset as the observable of the microbiome’s current structure.

I claimed that pragmatic information was our best bet at quantifying information in complex contexts; of course, it is still far from straightforward. If we define two probability distributions,  $P$  and  $Q$ , which represent the behavior of Bob before and after receiving the signal, respectively, we might define a measure of pragmatic information as

$$I_{pragmatics} = D_{KL}(P \parallel Q) = \sum_{a \in A} P(A) \log_2 \frac{P(A)}{Q(A)}, \quad (3.4)$$

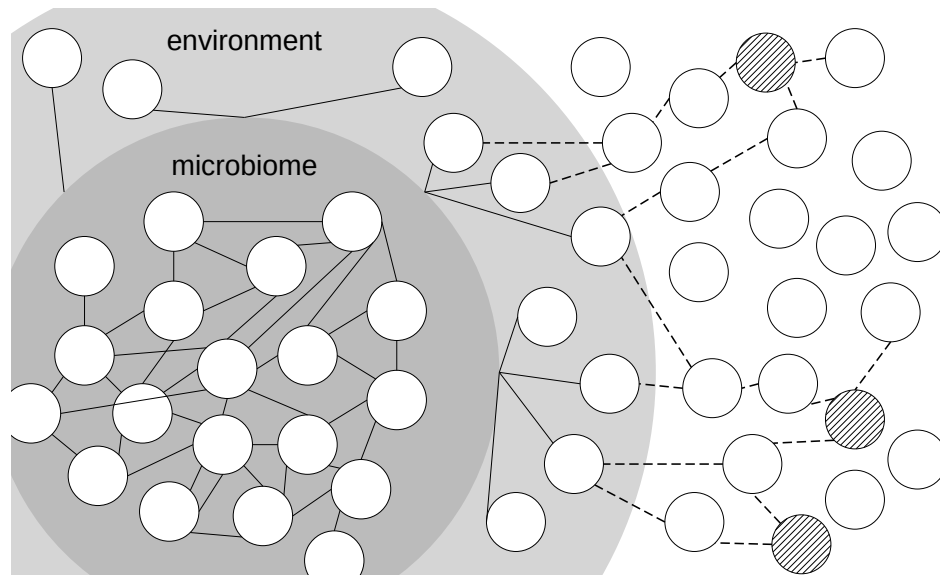
where  $A$  is the set of possible actions or distinct “pieces” of the behavior of Bob<sup>315,316</sup>. In other words, the pragmatic information of a signal on Bob is the Kullback-Leibler divergence between Bob’s behavior before ( $P$ ) and after ( $Q$ ) receiving a signal. While this might be close to the intuition behind pragmatic information, we cannot apply this formula to most study systems because it requires us to discretize actions and determine the probabilities of every single action of the system before and after receiving the message. If not previously known, one can only attain this data by repeatedly exposing Bob to the same signal and estimating the probability distributions from this experimental setting. However, this would evoke a case of circular logic hidden in the tripartite notion of information: On the one hand, the translation of syntactical information to pragmatic information is strongly dependent on the structure of Bob but, on the other hand, we measure pragmatic information by structural and behavioral changes. Thus, after receiving a message, Bob’s structure might change to the degree that repeating the message to Bob might lead to drastically different outcomes. After receiving the message, Bob might just not be comparable to Bob before receiving the message. This issue has been recognized in the literature as the problem of novelty and confirmation of messages<sup>107,167,311,312</sup>. However, this obstacle is ameliorated as long as we can reasonably assume that the changes that the pragmatic information represents impact the state of Bob in only a minor way (i.e., a change in microstates; see assumption II in section 1.4.2). Of course, this assumption will not hold if the signal induces a regime shift, a tipping point, or a change to a different attractor in Bob.

### 3.3.4 THE SENSOR AND ITS UMWELT

Maybe the most striking feature of the tripartite notion of information is that all three, syntactic, semantic, and pragmatic information, are strongly dependent on Bob's structure. This is in stark contrast to the formulation of information provided by Shannon, in which the independence of the message from Alice and Bob is essential. In fact, most of the parts essential to the model of communication, such as the channel, the sender, and even Alice, can be omitted from a graphical model that describes semiotic information (compare figures 3.1 and 3.2). This difference allows us to move away from the abstract set-up that surrounds Bob to a sensor or biosensor in the "real world", which is populated by objects and processes that might act as a signal. Which of these signals reach the sensor as syntactic information depends on the sensor's structure, which, thus, acts as an input filter. Or, in other words, only those signals can act as information for the sensor that are registered by its receiver structure. The measuring device will be oblivious to everything else.

This distinction splits the world in two; there is the world-at-large, which might only indirectly influence the sensor, and the sensor's environment, which contains all those objects and processes that the sensor receives information from<sup>169,303</sup>. If the structure of the sensor selects the information the sensor receives, then this environment is (albeit not completely actively) constructed by and relative to the sensor<sup>267,268</sup>. This idea goes back to at least the year 1926 when Jakob von Uexküll developed the concept of the "*Umwelt*" (akin to a direct environment), which is centrally defined by the "*Innenwelt*" (inner world) of the sensor (or, in his case, the organism). In Kantian fashion, this leads to the construction of a "*eine Welt für sich*" (a world of its own) that is incommensurable with the "world-in-itself"<sup>310</sup>. Inversely, this means that only objects and processes in the *Umwelt* of the sensor can act as the source of information – syntactic, semantic, or pragmatic – for the sensor. Everything outside of the *Umwelt* can only indirectly act on or interact with the sensor.

Although this vocabulary – and the theory underlying it – might seem too pompous when talking about anthropogenic sensors, it might serve as an excellent example of the use of the term *Umwelt*. These measuring devices are constructed in such a way as to sense only a particular set of properties of their *Umwelt*: In the environment of the thermometer, the kitchen scale, and the Geiger counter, there is only temperature, weight, and ionizing radiation, respectively. Furthermore, the objects in their environment are rather isolated – the only process affecting the weight



**Figure 3.3:** A graphical model for coupling of the microbiome. Each of the nodes represents one parameter in an ecosystem; some of them belong to the microbiome (and are, e.g., OTU numbers), some to the microbiome's direct environment and others to the world-at-large. If the interactions between the microorganisms (represented by edges that connect nodes in the microbiome) are taken into account by models that estimate coupling, the parameters in the microbiome's environment can be seen to interact with the microbiome as a whole and might be the domain of pragmatic information. Contrast this with parameters outside of the environment, which only interact with the microbiome indirectly. The actual interaction pathway of those parameters that are measured (represented by dashed lines and hatched nodes, respectively) are not trivial: The measured parameters might be influenced by other parameters or might influence multiple parameters in the microbiome's environment. Thus, these pathways will not be known, hard to identify, and not indicate direct correlation let alone causation. Instead, as argued in the main text, the relationship between these parameters and the microbiome is one of coupling.

measured with a kitchen scale, for example, is the user putting something on the scale. In fact, all other interactions of the sensors would be judged as malfunctions. As such, these measuring devices are trivial sensors.

In contrast to this, biosensors useful for biomonitoring inhabit a complex world populated by a high number of objects interacting with each other, but doing so in a comparatively sparse manner: Not every parameter interacts with all the other parameters. Each of the organisms that constitute the microbiome inhabits a *Umwelt* rich with objects and processes, that are, in turn, connected to many other objects and processes. When we are ready to traverse a sufficiently large number of interactions or information transfers, we will, eventually, reach any parameter of choice in the world-at-large (see figure 3.3 for a graphical representation).

For those parameters that are readily estimated from the changes observed in the biosensor, what I have stated above can be ignored safely. Such interactions are most probably in the domain of pragmatic information because the environmental parameter involved is most likely in the biosensor's environment. One would, furthermore, be tempted to disregard all attempts of approximating an environmental parameter that lead to low predictability as useless. If I were to adopt this position, I would need to conclude that the covariation measured between the lake microbiome and the environmental parameters presented in figure 3b of paper II (section 2.2) and figure 2 of paper III (section 2.3) only testify for the bad performance of the lake microbiome as a biosensor.

However, it was clear from the start that most parameters analyzed in the publications included herein are not in the microbiome's environment but might nevertheless affect it. Take the land cover categories analyzed in paper III (section 2.3): Of course, the microbiome is not able to sense these directly, and thus, it is clear that they do not belong in its environment. Nevertheless, it is just as obvious that changes in land cover impact the lake microbiome via physico-chemical variables that belong to the microbiome's environment. The same might be the case for physico-chemical parameters as, e.g., the concentration of calcium (which is one of the parameters analysed in paper II, section 2.3): While we measured its total concentration in the water samples, the microbiome might sense it on a different spatial scale or only when it is in a specific chemical configuration. We need to accept that we, through our measuring devices, have constructed a *Umwelt* that might have no intersection with the microbiome's *Umwelt* – or, in the words of Jakob von Uexküll: “If an observer has before him an animal whose world he wishes to investigate, he must first and foremost realise that the indications that make up the world of this other creature are his own, and do not originate from the marksigns of the animal's subject, which he cannot know in the least.”<sup>309</sup>

Using a biosensor is, thus, less like using a kitchen scale and more like a case of second-order observation: We are observing changes in the microbiome and its covariation with an environmental parameter we are observing while assuming that the microbiome is observing the changes in the same parameter, albeit indirectly. In other words: We are trying to look through the – metaphorical, mind you – eyes of the microbiome, squinting to see the parameter of interest, which lies across the horizon of the microbiome's direct environment. Second-order observation is an object of study for a wide range of scientific disciplines whose insights might, if translated

correctly, lead to theoretical advancements in the study of environmental microbiomes<sup>35,90,191,192,216,303</sup>.

One example of a concept important in second-order observation is structural coupling. In the work of the sociologist Niklas Luhmann, this term (originally “*strukturelle Kopplung*”) denotes the apparent coordination between complex systems that are, by definition, organisationally closed and, by that, functionally isolated from each other<sup>191,192</sup>. In the context of microbiome-based biomonitoring, the concept of coupling might serve as a quantitative measure of the relationship between parameters outside of the microbiome’s environment and the microbiome itself. As such, it would reflect the intuition that the microbiome might be more responsive to changes in some environmental parameters than in others. Or, in more detail: The changes in the environmental parameter we measured lead to and are induced by changes in other parameters, which, in turn, interact with other environmental parameters, and so on, until parameters are affected that are part of the microbiome’s environment, which, finally, transmit pragmatic information to the microbiome. Therefore, we have now found an answer to the initial question: The covariation framework estimates the degree of coupling between the microbiome and the parameter in question.

Now, what have we won with this insight? Structural coupling, as defined here, adheres to the limitations of statistical methods in complex systems set out in section 1.2.1: Coupling neither points to direct correlative, nor causal relationships and is not directional. Instead, it describes an apparent relationship, which might intuitively be interpreted as a distance measure, stating the functional distance between processes, objects, systems or assemblages that do not lie inside each other’s environments. For the analysis of environmental microbiomes, coupling can serve a few purposes: First, coupling allows us to gain initial insights into the importance of environmental parameters for the microbiome, potentially guiding further mechanistic studies. Second, it enables comparisons of parameters, e.g., located at different levels of hierarchy or belong to disparate domains. For many other methodologies, the comparison of land cover, physico-chemical, socio-economic, and weather parameters would not be possible. As I will further discuss in section 3.4, the results of comparing degrees of coupling for disparate sets of parameters are, in essence, structural and qualitative, but meaningful nevertheless.

To conclude this section of the discussion, I want to review the line of argumentation laid out here: In microbiome-based biomonitoring, the microbiome acts as something akin to a sensor. While one might want to explain its functioning as a process of information transference, the



classical notion of information does not fit here. Instead, the tripartite notion of information derived from semiotics seems more appropriate, but applying it to complex systems leads to a separation of the world into the environment functionally surrounding the object of study and the world-at-large. Finally, by adopting a concept from sociology, the covariation between the microbiome and a parameter outside of the microbiome's environment denotes the coupling of these two.

### 3.4 THE COVARIATION FRAMEWORK: COUPLING AND COMPLEXITY

This section will return to the covariation framework to contextualize it with what has been discussed in the previous section and point out how to develop it further. To that end, let me reiterate that, formally, by defining the microbiome as an assemblage that is a part of the ecosystem (section 1.2.3), we create a distinction between two complex sets of objects: the microbiome itself and the other environmental parameters, both of which are complex albeit rather assemblages than systems. Thus, both the microbiome as well as the environmental parameters exhibit the statistical obstacles of complexity introduced in section 1.2.1, including the apparent non-linearity of interactions as well as the inability to control for confounders. The covariation framework is constructed to handle both domains of complexity as follows.

The central idea behind the covariation framework is to interpret the output of a machine learning model trained to approximate an environmental parameter of interest as a mapping of the whole microbiome's community structure to a single dimension in the space of the environmental parameter. This procedure achieves two purposes: First, using machine learning models that can learn non-linear relationships from sparse, dependent, and high-dimensional datasets, an implicit model of the microbiome's response to changes in the environmental parameter is generated. This way, the microbiome's complexity is "abstracted away" in the model. Second, in contrast to the whole OTU table, the model's output can readily be compared to the environmental parameter in question. Given successful models, the output represents the variation of the microbiome relevant to the variation of the environmental parameter, reduced to a single dimension. Realistically, the model will underestimate the real coupling because (i) it is improbable that the model will be able to learn all aspects of the relationship between the model and the environmental parameter, and (ii) faulty overestimation (i.e., over-fitting of the model) is ruled out by evaluating the model via cross-validation.

As shown in figure 2 of publication II (section 2.2), different types of machine learning models lead to considerably (as well as significantly) different estimates of coupling, which confirms the intuition that the differences in the internals of the models have an important effect on the performance of the models. As could be expected from their description in section 1.3.4.3, ensembles of decision trees outperform other models. Based on these results, we might expect artificial neural networks (ANNs) to perform this task even better: By combining multiple neurons, which implement simple but tuneable, non-linear functions, in parallel and in series, they can be seen as universal approximators given enough training data and training time<sup>145,177</sup>. However, in this thesis, ANNs were not considered as they are notorious for requiring datasets with high numbers of samples for training successfully. Another alternative to Random Forest models comes in the form of equation-free modeling of complex systems by embedding their dynamical changes over time in attractor-space<sup>329,328</sup>. In a line of publications, such an approach has not only allowed for the identification of causal relationships between non-linearly interacting parameters but also the prediction of algal blooms in the coastal waters close to La Jolla, California, which were thought to be largely stochastic before<sup>202,291</sup>. To what extent this can be used in biomonitoring-like settings is unclear; it was not suitable for the datasets analyzed here because it requires time-series data.

To handle the complexity of the environment, the covariation framework separately creates a single model for each environmental parameter of interest. I chose this approach since a single, general model of the microbiome's dynamics would probably underestimate the complexity at play<sup>58</sup>. In a seminal paper, Richard Levins argued that ecological models cannot, at the same time, show high degrees of generality, realism, and precision toward understanding and predicting, in part because we can only partially specify the ecosystems they are to model<sup>179,181,220</sup>. Along these lines, the covariation framework initially sacrifices generality by analyzing separate environmental parameters separately but partially regains it by implementing a form of model pluralism, i.e., the parallel use of multiple models that describe different aspects of the study object<sup>82,305</sup>.

However, by that, the covariation framework cannot directly take into account the dependencies between environmental parameters or the microbiome's response to separate environmental parameters. To do so, one could implement so-called classifier or regressor chains<sup>129,203,247</sup>. These ensemble models consist of a series of machine learning models (as "base learners"), the first of which approximates an environmental parameter  $a$  based on the microbial community composi-

tion, the second of which uses the microbiome and the predicted value of  $a$ ,  $a'$  as input variables to predict  $b$ , and so on, until all target variables or environmental parameters are assigned to a model. As part of my research, I have attempted to apply model chains to the datasets studied here, but these turned out to underperform when compared to single models, most probably because of error-propagation in model chains<sup>276</sup>. Along the same lines, densely connected ANNs should be able to model dependencies between multiple input and multiple target variables. While this constitutes another reason to consider ANNs for further studies of environmental microbiomes, the high data requirements remains.

A further issue of ANNs, as well as chain models, is their lack of intuitive interpretability. The microbiome is, because of its complex nature, akin to a black box. This is made even more pressing if we assume the tripartite notion of information section, as it asserts that all three aspects of information are context-dependent and self-reflexive 3.3.3)<sup>334</sup>. However, pragmatic information introduces a way to gain insights into the object's syntactic and semantic structure – even if these are merely qualitative insights and not fully quantitative results. In the covariation framework, this is achieved by equipping us with an estimate of coupling and the output of the IndVal function (section 1.3.3). Given these, there are two points of view one can take to analyze the microbiome's internal structure: One from the outside, looking into the microbiome, to then see how the microbiome responds, and one from the viewpoint of the microbiome, which leads to an idea how the environment of the microbiome is structured for the microbiome<sup>166</sup>. For both, applications are developed in publication III (section 2.3), which are derived from comparisons of the lists of bioindicators. Both can be motivated in analogy to probing the function of a measuring device by making it interact with a set of different objects or processes. While one often cannot retroactively generate a full diagram of the sensor's internal wiring this way, something like the latent structures inside the measuring device will nevertheless emerge. The “response map” mentioned in publication III is an example for such a qualitative description of the microbiome's internals.

Thus, by examining the lists of bioindicators generated with the covariation framework in the context of pragmatic information, we can ask what the microbiome can “see”, i.e., what is in its environment, or, more formally, what distinctions it makes<sup>7</sup>. Similarly, it might be just as interesting to see what the microbiome does not “see” because this also reflects its structure<sup>142</sup>. For example, a recent, large-scale study of microbiomes in geothermal springs found no signif-

icant influence of temperature on the microbial community structure if the temperature was below 70 °C<sup>241</sup>, which strongly indicates that the geothermal microbiome can only distinguish between temperatures higher than that. Put formally, the question behind this is the following: If “information is a difference that makes a difference”<sup>21</sup>, then we might ask “of what nature a system must be so as to make a difference make a difference”<sup>35</sup>. What microorganisms need to be absent from or present in the microbiome in order for the whole community not to be affected by changes in temperature below 70 °C? Or, stated in molecular terms: What proteins make up these bacteria so that they are insensitive to temperature changes? Or, in evolutionary terms: What kinds of environmental selection might bring forward this form of temperature blindness? Therefore, and to conclude, a paradigm involving a strong notion of pragmatic information enables the asking of novel questions about the microbiome and allows one to do so from a broad range of biological disciplines.

### 3.5 THE FUTURE OF THE MICROBIOME

To end this thesis, let me state what kinds of questions arise from the results and theory presented here. Most of the more fundamental issues this thesis wrestles with stem from the fact that the data I work with are not derived from well-controlled experimental settings. Instead, they were sampled from real-world ecosystems and are observational in nature and, therefore, more suitable for data-driven than a hypothesis-driven mode of doing science<sup>227</sup>. While one might tend to object to the approach chosen here and take the theoretical contortions taken in this thesis, especially those in section 3.3, as proof for this, I hold that my approach has its merits. For one, the invasiveness of most experimental approaches makes them unsuitable for the study of ecosystems<sup>249,262,295</sup>. Furthermore, and as suggested in section 1.1, it might be just impossible to study real-world ecosystems experimentally, and the experimental setting might, systematically, underestimate the complexity of the study object<sup>240,248,331</sup>.

I expect that most future insights will arise from applying cutting-edge computational methods such as those developed here to long time-series datasets of environmental microbiomes. In such a setting, the assumption that samples reflect the dynamics of one underlying system is most likely to hold. Furthermore, time-series data would facilitate incorporating temporal changes of environmental microbiomes into both theory and methodology. To further support the endeavor of studying microbiomes as complex systems, we might also need to develop experimen-

tally accessible model systems for non-experimentally accessible environmental microbiomes<sup>39</sup>. While they come with a lot of additional issues, host-associated microbiomes might fill this void. They can be considered “microbiomes on a leash” and might harbor dynamics analogous environmental ones<sup>94,152</sup>. Along the same lines, artificially producible and simple microbiomes such as the one encountered in sourdough or microcosms might be models for environmental microbiomes as they exhibit even lower degrees of complexity than their host-associated counterparts<sup>84,172,206,314</sup>. By being imperfect experimental systems in that, e.g., not all parameters of their dynamics are fully controllable, they resemble environmental microbiomes more than, e.g., artificial microbial communities consisting of only a handful of well-known organisms.

After going through many current publications that study microbiomes, I cannot help to feel that it is also a lack of theory that makes the field stall in a descriptive phase<sup>28,76,205,216,283,327</sup>. One of the indicators of this is the overgrowth of nomenclatures of poorly defined terms and the number of reviews trying to domesticate the terminology, and with it, the study objects<sup>131,171,263,282</sup>. All other things being equal, it is imperative that we develop reliable mechanistic frameworks for how microbiomes act to support ecosystem health because the microscopic scale might be useful in averting the catastrophic effects on ecosystems currently caused by anthropogenic stressors<sup>313</sup>.

A theory of environmental microbiomes will require discourse between the disciplines of (molecular) microbiology, ecology, and computer science. I hope that the reader will appreciate that this thesis stands witness that such an interdisciplinary approach is not only possible but also instrumental in studying environmental microbiomes. However, I suspect that what will be necessary is the additional adaptation of theoretical results from the humanities and philosophy. Knowingly or unknowingly, these disciplines have, while describing the human condition, societies, and politics in non-experimental, real-world settings, worked on dynamics that we consider complex. In the course of that, these disciplines have produced consistent terminologies and theoretical methodologies that might prove helpful for studying environmental microbiomes. For example, take the theory of social systems developed for and applied to social systems by Niklas Luhmann. While I have only scratched the surface of his work, his abstract description of complex systems and the notion of coupling developed in his theory have proven very fruitful thinking about the work presented here. After all, is it too off-kilter to liken the microbiome to something like a society of microbes? A limitation of his work is that it lacks the ambition for quantitative analyses as far as I can tell. This gap might be filled by the “Free Energy Principle”, a

high-level mathematical description of complex systems as embodying minimal causal models of their environment in a Bayesian manner, use active inference to adapt to and modify their environment<sup>48,98,99,164,246</sup>. While not explicitly referencing Luhmannian theory, the literature on the Free Energy Principle makes use of similar concepts, including autopoiesis and a strong reading of *Umwelt*.

With all that being said, a highly diverse methodology and a plethora of literature to read makes it even more essential not to lose focus on the biology at hand. Otherwise, there is the danger of a headless methodology running wild – a danger far from new. Take, for example, the biting remarks that Richard Levins formulated in 1968 against the mathematical biology of his day: “[A]n octopus must move straight toward its prey; a transparent screen directly blocking its path is enough to thwart it. Chickens can make detours around obstacles as long as the goal is kept visible at all times. Dogs are able to turn their backs completely on the food in order to get around more difficult obstacles. And only mathematical biologists can turn their backs completely on their goal, wander off indefinitely, overcome obstacles, and smile. We are all painfully aware of the latter danger.”<sup>180</sup>

# References

- [1] K. Aggarwal (2003). Functional genomics and proteomics as a foundation for systems biology. *Briefings in Functional Genomics and Proteomics*, 2(3):175–184.
- [2] Craig R. Allen, David G. Angeler, Ahjond S. Garmestani, Lance H. Gunderson, and C. S. Holling (2014). Panarchy: Theory and application. *Ecosystems*, 17(4):578–589.
- [3] Timothy Allen and Thomas Hoekstra (2015). *Toward a Unified Ecology*. Columbia University Press.
- [4] Timothy F.H. Allen and E. Paul Wyleto (1983). A hierarchical model for the complexity of plant communities. *Journal of Theoretical Biology*, 101(4):529–540.
- [5] Carmen Astudillo-García, Syrie M. Hermans, Bryan Stevenson, Hannah L. Buckley, and Gavin Lear (2019). Microbial assemblages and bioindicators as proxies for ecosystem health status: potential and limitations. *Applied Microbiology and Biotechnology*, 103(16):6407–6421.
- [6] Harald Atmanspacher (1989). The aspect of information production in the process of observation. *Foundations of Physics*, 19:553–577.
- [7] Harald Atmanspacher (2007). A semiotic approach to complex systems. In *Aspects of Automatic Text Analysis*, pages 79–91. Springer Berlin Heidelberg.
- [8] Mariette Awad and Rahul Khanna (2015). Support vector regression. In *Efficient Learning Machines*, pages 67–80. Apress.
- [9] F. Azam, T. Fenchel, J.G. Field, J.S. Gray, L.A. Meyer-Reil, and F. Thingstad (1983). The ecological role of water-column microbes in the sea. *Marine Ecology Progress Series*, 10:257–263.

- [10] Philippe De Backer, Danny De Waele, and Linda Van Speybroeck (2009). Ins and outs of systems biology vis-à-vis molecular biology: Continuation or clear cut? *Acta Biotheoretica*, 58(1):15–49.
- [11] Mohammad Bahram, Falk Hildebrand, Sofia K. Forslund, Jennifer L. Anderson, Nadejda A. Soudzilovskaia, Peter M. Bodegom, Johan Bengtsson-Palme, Sten Anslan, Luis Pedro Coelho, Helery Harend, Jaime Huerta-Cepas, Marnix H. Medema, Mia R. Maltz, Sunil Mundra, Pål Axel Olsson, Mari Pent, Sergei P. Olme, Shinichi Sunagawa, Martin Ryberg, Leho Tedersoo, and Peer Bork (2018). Structure and function of the global topsoil microbiome. *Nature*, 560(7717):233–237.
- [12] J. Mark Baldwin (1896). A new factor in evolution. *The American Naturalist*, 30(354):441–451.
- [13] Samiran Banerjee, Klaus Schlaeppli, and Marcel G. A. van der Heijden (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology*, 16:567–576.
- [14] Samiran Banerjee, Klaus Schlaeppli, and Marcel G. A. van der Heijden (2018). Reply to ‘Can we predict microbial keystones?’. *Nature Reviews Microbiology*, 17(3):194–194.
- [15] Yinon M. Bar-On, Rob Phillips, and Ron Milo (2018). The biomass distribution on earth. *Proceedings of the National Academy of Sciences*, 115(25):6506–6511.
- [16] Yaneer Bar-Yam (2004). A mathematical theory of strong emergence using multiscale variety. *Complexity*, 9(6):15–24.
- [17] Marcello Barbieri (2012). The paradigms of biology. *Biosemiotics*, 6(1):33–59.
- [18] David Bass and Jens Boenigk (2011). Everything is everywhere: a twenty-first century de-/reconstruction with respect to protists. In Diego Fontaneto, editor, *Biogeography of Microscopic Organisms: Is Everything Small Everywhere?*, pages 88–110. Cambridge University Press.
- [19] Gregory Bateson (1963). The role of somatic change in evolution. *Evolution*, 17(4):529–539.



- [20] Gregory Bateson (1967). Cybernetic explanation. *American Behavioral Scientist*, 10(8):29–29.
- [21] Gregory Bateson (1970). Form, substance, and difference. *General Semantics*, 37.
- [22] B.E. Beisner, D.T. Haydon, and K. Cuddington (2003). Alternative stable states in ecology. *Frontiers in Ecology and the Environment*, 1(7):376–382.
- [23] Lyria Berdjeb, Jean François Ghiglione, Isabelle Domaizon, and Stéphan Jacquet (2010). A 2-year assessment of the main environmental factors driving the free-living bacterial community structure in Lake Bourget (France). *Microbial Ecology*, 61(4):941–954.
- [24] Gabriele Berg, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, Luca Cocolin, Kellye Eversole, Gema Herrero Corral, Maria Kazou, Linda Kinkel, Lene Lange, Nelson Lima, Alexander Loy, James A. Macklin, Emmanuelle Maguin, Tim Mauchline, Ryan McClure, Birgit Mitter, Matthew Ryan, Inga Sarand, Hauke Smidt, Bettina Schelkle, Hugo Roume, G. Seghal Kiran, Joseph Selvin, Rafael Soares Correa de Souza, Leo van Overbeek, Brajesh K. Singh, Michael Wagner, Aaron Walsh, Angela Sessitsch, and Michael Schlöter (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, 8(1).
- [25] Mark H. Bickhard and Donald T Campbell (2000). Emergence.
- [26] Sebastian Birk, Wendy Bonne, Angel Borja, Sandra Brucet, Anne Courrat, Sandra Poikane, Angelo Solimini, Wouter van de Bund, Nikolaos Zampoukas, and Daniel Hering (2012). Three hundred ways to assess Europe’s surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. *Ecological Indicators*, 18:31–41.
- [27] Christopher M. Bishop (2006). *Pattern Recognition And Machine Learning*. Springer Verlag.
- [28] Mariano Bizzarri, Douglas E. Brash, James Briscoe, Verônica A. Grieneisen, Claudio D. Stern, and Michael Levin (2019). A call for a better understanding of causation in cell biology. *Nature Reviews Molecular Cell Biology*, 20(5):261–262.

- [29] Christina Bock, Manfred Jensen, Dominik Forster, Sabina Marks, Julia Nuy, Roland Psenner, Daniela Beisser, and Jens Boenigk (2020). Factors shaping community patterns of protists and bacteria on a european scale. *Environmental Microbiology*, 22(6):2243–2260.
- [30] Jens Boenigk, Sabina Wodniok, Christina Bock, Daniela Beisser, Christopher Hempel, Lars Grossmann, Anja Lange, and Manfred Jensen (2018). Geographic distance and mountain ranges structure freshwater protist communities on a european scale. *Metabarcoding and Metagenomics*, 2:e21519.
- [31] Angel Borja, Mike Elliott, Jesper H. Andersen, Ana C. Cardoso, Jacob Carstensen, João G. Ferreira, Anna-Stiina Heiskanen, João C. Marques, João M. Neto, Heliana Teixeira, Laura Uusitalo, María C. Uyarra, and Nikolaos Zampoukas (2013). Good environmental status of marine ecosystems: What is it and how do we know when we have attained it? *Marine Pollution Bulletin*, 76(1-2):16–27.
- [32] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory (COLT '92)*, page 144–152. ACM Press.
- [33] Leo Breiman (2001). Random forests. *Machine Learning*, 45(1):5–32.
- [34] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen (1984). *Classification And Regression Trees*. Chapman and Hall/CRC.
- [35] Søren Brier (1996). From second-order cybernetics to cybersemiotics: A semiotic re-entry into the second-order cybernetics of Heinz von Foerster. *Systems Research*, 13(3):229–244.
- [36] Daniel S. Brooks (2019). A new look at ‘levels of organization’ in biology. *Erkenntnis*.
- [37] Yu. S. Bukin, Yu. P. Galachyants, I. V. Morozov, S. V. Bukin, A. S. Zakharenko, and T. I. Zemskaya (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, 6(1).
- [38] Van Butsic, David J. Lewis, Volker C. Radeloff, Matthias Baumann, and Tobias Kuemmerle (2017). Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic and Applied Ecology*, 19:1–10.

- [39] Marc W. Cadotte, James A. Drake, and Tadashi Fukami (2005). Constructing nature: Laboratory models as necessary tools for investigating complex ecological communities. *Advances in Ecological Research*, 37:333–353.
- [40] Benjamin J. Callahan, Paul J. McMurdie, and Susan P. Holmes (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12):2639–2643.
- [41] Benjamin J. Callahan, Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes (2016). DADA2: high-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7):581–583.
- [42] S. Chaffron, H. Rehrauer, J. Pernthaler, and C. von Mering (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947–959.
- [43] Tianqi Chen and Carlos Guestrin (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.
- [44] Ashley R. Coenen and Joshua S. Weitz (2018). Limitations of correlation-based inference in complex virus-microbe communities. *mSystems*, 3(4).
- [45] Patrick Collard (1976). *The Development of Microbiology*. Cambridge University Press.
- [46] John Collier (1988). Supervenience and reduction in biological hierarchies. *Canadian Journal of Philosophy Supplementary Volume*, 14:209–234.
- [47] R. R. Colwell (1997). Microbial diversity: the importance of exploration and conservation. *Journal of Industrial Microbiology and Biotechnology*, 18(5):302–307.
- [48] Axel Constant, Maxwell J. D. Ramstead, Samuel P. L. Veissière, John O. Campbell, and Karl J. Friston (2018). A variational approach to niche construction. *Journal of The Royal Society Interface*, 15(141):20170685.
- [49] Tristan Cordier, Laura Alonso-Sáez, Laure Apothéloz-Perret-Gentil, Eva Aylagas, David A. Bohan, Agnès Bouchez, Anthony Chariton, Simon Creer, Larissa Frühe, François Keck, Nigel Keeley, Olivier Laroche, Florian Leese, Xavier Pochon, Thorsten Stoeck, Jan Pawlowski, and Anders Lanzén (2020). Ecosystems monitoring powered by

environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*.

- [50] Tristan Cordier, Dominik Forster, Yoann Dufresne, Catarina I. M. Martins, Thorsten Stoeck, and Jan Pawlowski (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, 18(6):1381–1391.
- [51] Tristan Cordier, Anders Lanzén, Laure Apothéloz-Perret-Gentil, Thorsten Stoeck, and Jan Pawlowski (2019). Embracing environmental genomics and machine learning for routine biomonitoring. *Trends in Microbiology*, 27(5):387–397.
- [52] Sam Cowling (2017). Resemblance. *Philosophy Compass*, 12(4):e12401.
- [53] Carl F. Craver and William Bechtel (2006). Top-down causation without top-down causes. *Biology & Philosophy*, 22(4):547–563.
- [54] D.H. Cropley (1998). Towards formulating a semiotic theory of measurement information – part 1. *Measurement*, 24(4):237–248.
- [55] D.H. Cropley (1998). Towards formulating a semiotic theory of measurement information – part 2. *Measurement*, 24(4):249–262.
- [56] V. Csányi and G. Kampis (1985). Autogenesis: The evolution of replicative systems. *Journal of Theoretical Biology*, 114(2):303–321.
- [57] Jaime F. Cárdenas-García and Timothy Ireland (2019). The fundamental problem of the science of information. *Biosemitotics*, 12(2):213–244.
- [58] Virginia H. Dale and Suzanne C. Beyeler (2001). Challenges in the development and use of ecological indicators. *Ecological Indicators*, 1(1):3–10.
- [59] Miquel De Cáceres and Pierre Legendre (2009). Associations between species and groups of sites: indices and statistical inference.
- [60] Miquel De Cáceres, Pierre Legendre, and Marco Moretti (2010). Improving indicator species analysis by combining groups of sites. *Oikos*, 119(10):1674–1684.
- [61] J. de Haan (2006). How emergence arises. *Ecological Complexity*, 3(4):291–301.

- [62] Henk W. de Regt (2017). *Understanding Scientific Understanding*. Oxford University Press.
- [63] Rutger de Wit and Thierry Bouvier (2006). 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology*, 8(4):755–758.
- [64] Glenn De'ath (2002). Multivariate regression trees:: A new technique for modeling species-environment relationships. *Ecology*, 83(4):1105–1117.
- [65] Glenn De'ath (2007). Boosted trees for ecological modeling and prediction. *Ecology*, 88(1):243–251.
- [66] Glenn De'ath and Katharina E. Fabricius (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192.
- [67] Edward S. Deevey (1951). Life in the depths of a pond. *Scientific American*, 185.
- [68] Manuel Delgado-Baquerizo, Angela M. Oliverio, Tess E. Brewer, Alberto Benavent-González, David J. Eldridge, Richard D. Bardgett, Fernando T. Maestre, Brajesh K. Singh, and Noah Fierer (2018). A global atlas of the dominant bacteria found in soil. *Science*, 359(6373):320–325.
- [69] C. Lisa Dent, Graeme S. Cumming, and Stephen R. Carpenter (2002). Multiple states in river and lake ecosystems. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1421):635–645.
- [70] David Deutsch and Chiara Marletto (2015). Constructor theory of information. *Proceedings of the Royal Society A*, 471(2174):20140540.
- [71] Tom Dietterich (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, 27(3):326–327.
- [72] Kathryn M. Docherty and Jessica L. M. Gutknecht (2011). The role of environmental microorganisms in ecosystem responses to global change: current state of research and future outlooks. *Biogeochemistry*, 109(1-3):1–6.

- [73] Pedro Domingos (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78.
- [74] Harris Drucker, Chris J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik (1996). Support vector regression machines. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press.
- [75] Marc Dufrêne and Pierre Legendre (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs*, 67(3):345–366.
- [76] John Dupré and Maureen A. O’Malley (2007). Metagenomics and biological ontology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 38(4):834–846.
- [77] Frank N Egerton (2002). A history of the ecological sciences, part 6: arabic language science: origins and zoological writings. *Bulletin of the Ecological Society of America*, 83.
- [78] Frank N. Egerton (2007). Understanding food chains and food webs, 1700–1970. *Bulletin of the Ecological Society of America*, 88(1):50–69.
- [79] Markus Eichhorn (2016). *Natural Systems - The Organisation Of Life*. John Wiley & Sons.
- [80] Manfred Eigen (1971). Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58(10):465–523.
- [81] Claus Emmeche (1997). Autopoietic systems, replicators, and the search for a meaningful biologic definition of life. *Ultimate Reality and Meaning*, 20(4):244–264.
- [82] Claus Emmeche (1998). Defining life as a semiotic phenomenon. *Cybernetics & Human Knowing*, 5(1):3–17.
- [83] J. Engelberg and L. L. Boyarsky (1979). The noncybernetic nature of ecosystems. *The American Naturalist*, 114(3):317–324.
- [84] Danilo Ercolini, Erica Pontonio, Francesca De Filippis, Fabio Minervini, Antonietta La Storia, Marco Gobbetti, and Raffaella Di Cagno (2013). Microbial ecology dynamics dur-

- ing rye and wheat sourdough preparation. *Applied and Environmental Microbiology*, 79(24):7827–7836.
- [85] Markus I. Eronen (2014). Levels of organization: a deflationary account. *Biology & Philosophy*, 30(1):39–58.
  - [86] Ferric C. Fang and Arturo Casadevall (2011). Reductionistic and holistic science. *Infection and Immunity*, 79(4):1401–1404.
  - [87] Brian D. Fath and Bernard C. Patten (1999). Review of the foundations of network environ analysis. *Ecosystems*, 2(2):167–179.
  - [88] Brian D. Fath, Ursula M. Scharler, Robert E. Ulanowicz, and Bruce Hannon (2007). Ecological network analysis: network construction. *Ecological Modelling*, 208(1):49–55.
  - [89] James K. Feibleman (1954). Theory of integrative levels. *The British Journal for the Philosophy of Science*, 5(17):59–66.
  - [90] Bernard Feltz, Marc Crommelinck, and Philippe Goujon (2006). *Self-Organization And Emergence In Life Sciences*. Springer Science & Business Media.
  - [91] John T. Finn (1976). Measures of ecosystem structure and function derived from analysis of flows. *Journal of Theoretical Biology*, 56(2):363–380.
  - [92] Jo Foden, Stuart I. Rogers, and Andrew P. Jones (2008). A critical review of approaches to aquatic environmental assessment. *Marine Pollution Bulletin*, 56(11):1825–1833.
  - [93] Dominik Forster, Zhishuai Qu, Gianna Pitsch, Estelle P. Bruni, Barbara Kammerlander, Thomas Pröschold, Bettina Sonntag, Thomas Posch, and Thorsten Stoeck (2021). Lake ecosystem robustness and resilience inferred from a climate-stressed protistan plankton network. *Microorganisms*, 9(3):549.
  - [94] Kevin R. Foster, Jonas Schluter, Katharine Z. Coyte, and Seth Rakoff-Nahoum (2017). The evolution of the host microbiome as an ecosystem on a leash. *Nature*, 548(7665):43–51.
  - [95] Pedro R. Frade, Bettina Glasl, Samuel A. Matthews, Camille Mellin, Ester A. Serrão, Kennedy Wolfe, Peter J. Mumby, Nicole S. Webster, and David G. Bourne (2020). Spatial

- patterns of microbial communities across surface waters of the great barrier reef. *Communications Biology*, 3(442).
- [96] Jerome H. Friedman (2001). Greedy function approximation: A gradient boosting. *The Annals of Statistics*, 29(5):1189–1232.
  - [97] Jerome H. Friedman (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.
  - [98] Karl Friston (2012). A free energy principle for biological systems. *Entropy*, 14(11):2100–2121.
  - [99] Karl Friston (2013). Life as we know it. *Journal of The Royal Society Interface*, 10(86):20130475.
  - [100] Karl Friston, Christopher Thornton, and Andy Clark (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, 3.
  - [101] Larissa Frühe, Tristan Cordier, Verena Dully, Hans-Werner Breiner, Guillaume Lentendu, Jan Pawlowski, Catarina Martins, Thomas A. Wilding, and Thorsten Stoeck (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology*.
  - [102] Jed A. Fuhrman, Jacob A. Cram, and David M. Needham (2015). Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, 13(3):133–146.
  - [103] Natalia García-García, Javier Tamames, Alexandra M. Linz, Carlos Pedrós-Alió, and Fernando Puente-Sánchez (2019). Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. *The ISME Journal*, 13:2969–2983.
  - [104] Ahjond S. Garmestani, Craig R. Allen, and Lance Gunderson (2009). Panarchy: Discontinuities reveal similarities in the dynamic system structure of ecological and social systems. *Ecology and Society*, 14(1).
  - [105] Jeff Gauthier, Antony T Vincent, Steve J Charette, and Nicolas Derome (2018). A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6):1981–1996.



- [106] Hui Ge, Albertha J.M Walhout, and Marc Vidal (2003). Integrating ‘omic’ information: a bridge between genomics and systems biology. *Trends in Genetics*, 19(10):551–560.
- [107] Dieter Gernert (2006). Pragmatic information: Historical exposition and general overview. *Mind and Matter*, 4(2):141–167.
- [108] Carlos Gershenson and Francis Heylighen (2003). When can we call a system self-organizing? In *Advances in Artificial Life. ECAL 2003*, pages 606–614. Springer Berlin Heidelberg.
- [109] Carlos Gershenson and Francis Heylighen (2005). How can we think the complex? *Managing organizational complexity: philosophy, theory and application*, 3:47–62.
- [110] Sean M. Gibbons, Claire Duvall, and Eric J. Alm (2018). Correcting for batch effects in case-control microbiome studies. *PLOS Computational Biology*, 14(4):e1006102.
- [111] Jack A. Gilbert, Joshua A. Steele, J. Gregory Caporaso, Lars Steinbrück, Jens Reeder, Ben Temperton, Susan Huse, Alice C. McHardy, Rob Knight, Ian Joint, Paul Somerfield, Jed A. Fuhrman, and Dawn Field (2011). Defining seasonal marine microbial community dynamics. *The ISME Journal*, 6(2):298–308.
- [112] Bettina Glasl, David G. Bourne, Pedro Frade, Torsten Thomas, Britta Schaffelke, and Nicole S. Webster (2019). Microbial predictors of environmental perturbations in coral reef ecosystems. *Microbiome*, 7(94).
- [113] James Gleick (1997). *Chaos - Making A New Science*. Random House.
- [114] Didier Gonze, Katharine Z. Coyte, Leo Lahti, and Karoline Faust (2018). Microbial communities as dynamical systems. *Current Opinion in Microbiology*, 44:41–49.
- [115] Emily B. Graham, Joseph E. Knelman, Andreas Schindlbacher, Steven Siciliano, Marc Breulmann, Anthony Yannarell, J. M. Beman, Guy Abell, Laurent Philippot, James Prosser, Arnaud Foulquier, Jorge C. Yuste, Helen C. Glanville, Davey L. Jones, Roey Angel, Janne Salminen, Ryan J. Newton, Helmut Bürgmann, Lachlan J. Ingram, Ute Hamer, Henri M. P. Siljanen, Krista Peltoniemi, Karin Potthast, Lluís Bañeras, Martin Hartmann, Samiran Banerjee, Ri-Qing Yu, Geraldine Nogaro, Andreas Richter, Marianne Koranda, Sarah C. Castle, Marta Goberna, Bongkeun Song, Amitava Chatterjee, Olga C. Nunes,

- Ana R. Lopes, Yiping Cao, Aurore Kaisermann, Sara Hallin, Michael S. Strickland, Jordi Garcia-Pausas, Josep Barba, Hojeong Kang, Kazuo Isobe, Sokratis Papaspyrou, Roberta Pastorelli, Alessandra Lagomarsino, Eva S. Lindström, Nathan Basiliko, and Diana R. Nemergut (2016). Microbes as engines of ecosystem function: When does community structure enhance predictions of ecosystem processes? *Frontiers in Microbiology*, 7.
- [116] Sander Greenland and Hal Morgenstern (1989). Ecological bias, confounding, and effect modification. *International Journal of Epidemiology*, 18(1):269–274.
- [117] Volker Grimm (1996). A down-to-earth assessment of stability concepts in ecology: dreams, demands, and the real problems. *Senckenbergiana maritima*, 27.
- [118] Volker Grimm (1998). To be, or to be essentially the same: the ‘self-identity of ecological units’. *Trends in Ecology & Evolution*, 13(8):298–299.
- [119] Volker Grimm and Christian Wissel (1997). Babel, or the ecological stability discussions: an inventory and analysis of terminology and a guide for avoiding confusion. *Oecologia*, 109(3):323–334.
- [120] Lars Grossmann, Manfred Jensen, Dominik Heider, Steffen Jost, Edvard Glücksman, Hanna Hartikainen, Shazia S. Mahamdallie, Michelle Gardner, Daniel Hoffmann, David Bass, and Jens Boenigk (2016). Protistan community analysis: key findings of a large-scale molecular sampling. *The ISME Journal*, 10(9):2269–2279.
- [121] Lars Grossmann, Manfred Jensen, Ram V. Pandey, Steffen Jost, David Bass, Roland Psenner, and Jens Boenigk (2016). Molecular investigation of protistan diversity along an elevation transect of alpine lakes. *Aquatic Microbial Ecology*, 78(1):25–37.
- [122] Burton S. Guttman (1976). Commentary: Is ”levels of organization” a useful biological concept? *BioScience*, 26(2):112–113.
- [123] Ian Hacking (1983). *Representing and Intervening*. Cambridge University Press.
- [124] Efraim Halfon (1979). *Theoretical Systems Ecology - Advances And Case Studies*. Academic Press.
- [125] Maozhen Han, Melissa Dsouza, Chunyu Zhou, Hongjun Li, Junqian Zhang, Chaoyun Chen, Qi Yao, Chaofang Zhong, Hao Zhou, Jack A Gilbert, Zhi Wang, and Kang Ning

- (2019). Agricultural risk factors influence microbial ecology in honghu lake. *Genomics Proteomics & Bioinformatics*, 17(1):76–90.
- [126] Bruce Hannon (1973). The structure of ecosystems. *Journal of Theoretical Biology*, 41(3):535–546.
- [127] Lars-Anders Hansson, Jakob Brodersen, Ben B. Chapman, Mattias K. Ekvall, Anders Hargeby, Kaj Hulthén, Alice Nicolle, P. Anders Nilsson, Christian Skov, and Christer Brönmark (2013). A lake as a microcosm: reflections on developments in aquatic ecology. *Aquatic Ecology*, 47(2):125–135.
- [128] Alan Hastings, Carole L. Hom, Stephen Ellner, Peter Turchin, and H. Charles J. Godfray (1993). Chaos in ecology: Is mother nature a strange attractor? *Annual Review of Ecology and Systematics*, 24(1):1–33.
- [129] D. Heider, R. Senge, W. Cheng, and E. Hullermeier (2013). Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*, 29(16):1946–1952.
- [130] Dominik Heider, Christoph Bartenhagen, J. Nikolaj Dybowski, Sascha Hauke, Martin Pyka, and Daniel Hoffmann (2013). Unsupervised dimension reduction methods for protein sequence classification. In *Studies in Classification Data Analysis and Knowledge Organization*, pages 295–302. Springer Nature.
- [131] Ulrich Heink and Ingo Kowarik (2010). What are indicators? On the definition of indicators in ecology and environmental planning. *Ecological Indicators*, 10(3):584–593.
- [132] A.-S. Heiskanen, W. van de Bund, A.C. Cardoso, and P. Nöges (2004). Towards good ecological status of surface waters in Europe - interpretation and harmonisation of the concept. *Water Science and Technology*, 49(7):169–177.
- [133] Daniel Hering, Angel Borja, Jacob Carstensen, Laurence Carvalho, Mike Elliott, Christian K. Feld, Anna-Stiina Heiskanen, Richard K. Johnson, Jannicke Moe, and Didier Pont (2010). The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. *Science of The Total Environment*, 408(19):4007–4019.

- [134] Daniel Hering, Angel Borja, J. Iwan Jones, Didier Pont, Pieter Boets, Agnes Bouchez, Kat Bruce, Stina Drakare, Bernd Hänfling, Maria Kahlert, Florian Leese, Kristian Meissner, Patricia Mergen, Yorick Reyjol, Pedro Segurado, Alfried Vogler, and Martyn Kelly (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research*, 138:192–205.
- [135] Darrel Hess and Dennis G. Tasa (2016). *McKnight's Physical Geography - A Landscape Appreciation*. Prentice Hall.
- [136] Francis Heylighen (1999). Classical and non-classical representations in physics I. *Cybernetics and Systems*, 21(5):477–502.
- [137] M. Higashi and Thomas P. Burns (2009). *Theoretical Studies of Ecosystems - The Network Perspective*. Cambridge University Press.
- [138] Masahiko Higashi (1986). Extended input-output flow analysis of ecosystems. *Ecological Modelling*, 32(1-3):137–147.
- [139] Masahiko Higashi and Bernard C. Patten (1986). Further aspects of the analysis of indirect effects in ecosystems. *Ecological Modelling*, 31(1-4):69–77.
- [140] Masahiko Higashi and Bernard C. Patten (1989). Dominance of indirect causality in ecosystems. *The American Naturalist*, 133(2):288–302.
- [141] Hokuto Hirano and Kazuhiro Takemoto (2019). Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinformatics*, 20(329).
- [142] Jesper Hoffmeyer and Claus Emmeche (1991). Code-duality and the semiotics of nature. In *On Semiotic Modeling*, pages 117–166. De Gruyter Mouton.
- [143] Christopher Holder and Anand Gnanadesikan (2021). Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations? – A proof-of-concept study. *Biogeosciences*, 18(6):1941–1970.
- [144] Shoji Horie (1962). Morphometric features and the classification of all the lakes in japan. *Mem Coll Sci Univ Kyoto (B)*, 29.

- [145] Kurt Hornik (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.
- [146] Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas, and Jillian F. Banfield (2016). A new view of the tree of life. *Nat. Microbiol*, page 16048.
- [147] Luisa W. Hugerth, John Larsson, Johannes Alneberg, Markus V. Lindh, Catherine Legrand, Jarone Pinhassi, and Anders F. Andersson (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biology*, 16(1).
- [148] John Huss (2014). Methodology and ontology in microbiome research. *Biological Theory*, 9(4):392–400.
- [149] George Evelyn Hutchinson (1975). *A Treatise On Limnology: Geography, Physics, And Chemistry. Pt. 1. Geography And Physics Of Lakes*. Wiley.
- [150] Raban Iten, Tony Metger, Henrik Wilming, Lidia del Rio, and Renato Renner (2020). Discovering physical concepts with neural networks. *Physical Review Letters*, 124(1).
- [151] Kurt Jax (2006). Ecological units: Definitions and application. *The Quarterly Review of Biology*, 81(3):237–258.
- [152] Brian W. Ji, Ravi U. Sheth, Purushottam D. Dixit, Konstantine Tchourine, and Dennis Vitkup (2020). Macroecological dynamics of gut microbiota. *Nature Microbiology*, 5(5):768–775.
- [153] Lionel Johnson (1981). The thermodynamic origin of ecosystems. *Canadian Journal of Fisheries and Aquatic Sciences*, 38(5):571–590.
- [154] Adriane Clark Jones, K. David Hambright, and David A. Caron (2017). Ecological patterns among bacteria and microbial eukaryotes derived from network analyses in a low-salinity lake. *Microbial Ecology*, 75(4):1–13.
- [155] Sven E. Jørgensen, Søren Nors Nielsen, and Brian D. Fath (2016). Recent progress in systems ecology. *Ecological Modelling*, 319:112–118.

- [156] Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O’Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L. Mason, Karen L. Madsen, and Gane K.-S. Wong (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*, 7.
- [157] S.E. Jørgensen and S.N. Nielsen (2015). Hierarchical networks. *Ecological Modelling*, 295:59–65.
- [158] Immanuel Kant (1974). *Kritik der Reinen Vernunft*. Suhrkamp.
- [159] Battle Karimi, Pierre Alain Maron, Nicolas Chemidlin-Prevost Boure, Nadine Bernard, Daniel Gilbert, and Lionel Ranjard (2017). Microbial diversity and ecological networks as indicators of environmental quality. *Environmental Chemistry Letters*, 15(2):265–281.
- [160] Michael Kearns and Leslie Valiant (1994). Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM*, 41(1):67–95.
- [161] Steve Kelling, Wesley M. Hochachka, Daniel Fink, Mirek Riedewald, Rich Caruana, Grant Ballard, and Giles Hooker (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience*, 59(7):613–620.
- [162] Ryan S. King and Matthew E. Baker (2010). Considerations for analyzing ecological community thresholds in response to anthropogenic environmental gradients. *Journal of the North American Benthological Society*, 29(3):998–1008.
- [163] Ryan S. King and Curtis J. Richardson (2003). Integrating bioassessment and ecological risk assessment: An approach to developing numerical water-quality criteria. *Environmental Management*, 31(6):795–809.
- [164] Michael Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein (2018). The markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15(138):20170792.
- [165] A. N. Kolmogorov (1968). Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1-4):157–168.

- [166] Klaus Kornwachs (1996). Pragmatic information and system surface. In K. Kornwachs & K. Jacoby, editor, *Information. New questions to a multidisciplinary concept*, pages 163–185. Berlin: Akademie Verlag.
- [167] Klaus Kornwachs (1998). Pragmatic information and the emergence of meaning. In Delpo M. van de Vijver G., Salthe S.N., editor, *Evolutionary Systems*, pages 181–196. Springer Netherlands.
- [168] S. A. Kraemer, N. Barbosa da Costa, B. J. Shapiro, M. Fradette, Y. Huot, and D. A. Walsh (2020). A large-scale assessment of lakes reveals a pervasive signal of land use on bacterial communities. *The ISME Journal*, 14:3011–3023.
- [169] Kalevi Kull (2010). Ecosystems are made of semiotic bonds: Consortia, Umwelten, biophony and ecological codes. *Biosemiotics*, 3(3):347–357.
- [170] Bernd-Olaf Küppers (1996). The context-dependence of biological information. In K. Kornwachs & K. Jacoby, editor, *Information. New questions to a multidisciplinary concept*, pages 137–145. Berlin: Akademie Verlag.
- [171] James Ladyman, James Lambert, and Karoline Wiesner (2012). What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67.
- [172] Elizabeth A Landis, Angela M Oliverio, Erin A McKenney, Lauren M Nichols, Nicole Kfoury, Megan Biango-Daniels, Leonora K Shell, Anne A Madden, Lori Shapiro, Shravya Sakunala, Kinsey Drake, Albert Robbat, Matthew Booker, Robert R Dunn, Noah Fierer, and Benjamin E Wolfe (2021). The diversity and function of sourdough starter microbiomes. *eLife*, 10:e61644.
- [173] Peter B. Landres, Jared Verner, and Jack Ward Thomas (1988). Ecological uses of vertebrate indicator species: A critique. *Conservation Biology*, 2(4):316–328.
- [174] D. J. Lane, B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences*, 82(20):6955–9.

- [175] Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, and Victor Robles (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112.
- [176] Christian L Lauber, Kelly S Ramirez, Zach Aanderud, Jay Lennon, and Noah Fierer (2013). Temporal variability in soil microbial communities across land-use types. *The ISME Journal*, 7(8):1641–1650.
- [177] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton (2015). Deep learning. *Nature*, 521(7553):436–444.
- [178] Simon A. Levin (1998). Ecosystems and the biosphere as complex adaptive systems. *Ecosystems*, 1(5):431–436.
- [179] Richard Levins (1966). The strategy of model building in population biology. *American scientist*, 54.
- [180] Richard Levins (1968). Ecological engineering: Theory and technology. *The Quarterly Review of Biology*, 43(3):301–305.
- [181] Richard Levins (1974). Discussion paper: The qualitative analysis of partially specified systems. *Annals of the New York Academy of Sciences*, 231(1):123–138.
- [182] Richard Levins (1975). Evolution in communities near equilibrium. In Jared Diamond Martin Cody, editor, *Ecology and evolution of communities*, pages 16–50. Harvard University Press Cambridge, MA.
- [183] Richard Levins and Richard Lewontin (1980). Dialectics and reductionism in ecology. *Synthese*, 43(1):47–78.
- [184] Richard Levins and Richard Lewontin (1985). *The Dialectical Biologist*. Harvard University Press.
- [185] Raymond L. Lindeman (1942). The trophic-dynamic aspect of ecology. *Ecology*, 23(4):399–417.
- [186] Markus V. Lindh, Johanna Sjöstedt, Anders F. Andersson, Federico Baltar, Luisa W. Hugerth, Daniel Lundin, Saraladevi Muthusamy, Catherine Legrand, and Jarone Pinhassi



- (2015). Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling. *Environmental Microbiology*, 17(7):2459–2476.
- [187] Kenneth J. Locey and Jay T. Lennon (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975.
- [188] Hannah F. Löchel, Dominic Eger, Theodor Sperlea, and Dominik Heider (2019). Deep learning on chaos game representation for proteins. *Bioinformatics*, 36(1):272–279.
- [189] G. Longo, P.-A. Miquel, C. Sonnenschein, and A.M. Soto (2012). Is information a proper observable for biological organization? *Progress in Biophysics and Molecular Biology*, 109(3):108–114.
- [190] Catherine A. Lozupone, Micah Hamady, Scott T. Kelley, and Rob Knight (2007). Quantitative and qualitative  $\beta$  diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5):1576–1585.
- [191] Niklas Luhmann (1984). *Soziale Systeme: Grundriß einer allgemeinen Theorie*. Suhrkamp.
- [192] Niklas Luhmann (1990). Sthenographie. In *Beobachter – Konvergenz der Erkenntnistheorien?*, pages 119–137. Fink, München.
- [193] Niklas Luhmann (1993). Zeichen als Form. In Dirk Baecker, editor, *Probleme der Form*, pages 45–69. Suhrkamp Frankfurt am Main.
- [194] Denis H. Lynn and Guy L. Gilron (1992). A brief review of approaches using ciliated protists to assess aquatic ecosystem health. *Journal of Aquatic Ecosystem Health*, 1(4):263–270.
- [195] Donald MacCrimmon MacKay (1969). *Information, Mechanism and Meaning*. MIT Press (MA).
- [196] Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593.

- [197] Julian R. Marchesi and Jacques Ravel (2015). The vocabulary of microbiome research: a proposal. *Microbiome*, 3(1).
- [198] Ramon Massana and Ramiro Logares (2012). Eukaryotic versus prokaryotic marine picoplankton ecology. *Environmental Microbiology*, 15(5):1254–1261.
- [199] Humberto R. Maturana and Francisco J. Varela (1980). *Autopoiesis and Cognition*. D. Reidel Publishing Company.
- [200] Fulvio Mazzocchi (2008). Complexity in biology. *EMBO reports*, 9(1):10–14.
- [201] Melodie A McGeoch and Steven L Chown (1998). Scaling up the value of bioindicators. *Trends in Ecology & Evolution*, 13(2):46–47.
- [202] John A. McGowan, Ethan R. Deyle, Hao Ye, Melissa L. Carter, Charles T. Perretti, Kerri D. Seger, Alain Verneil, and George Sugihara (2017). Predicting coastal algal blooms in southern california. *Ecology*, 98(5):1419–1433.
- [203] Gabriella Melki, Alberto Cano, Vojislav Kecman, and Sebastián Ventura (2017). Multi-target support vector regression via correlation regressor chains. *Information Sciences*, 415-416:53–69.
- [204] M. D. Mesarović (1968). Systems theory and biology—view of a theoretician. In M. D. Mesarović, editor, *Systems Theory and Biology*, pages 59–87. Springer Berlin Heidelberg.
- [205] M.D. Mesarović, J.D. Keene, and S.N. Sreenath (2004). Search for organising principles: understanding in systems biology. *Systems Biology*, 1(1):19–27.
- [206] Fabio Minervini, Maria De Angelis, Raffaella Di Cagno, and Marco Gobbetti (2014). Ecological parameters influencing microbial diversity and stability of traditional sourdough. *International Journal of Food Microbiology*, 171:136–146.
- [207] Babak Momeni, Li Xie, and Wenying Shou (2017). Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *eLife*, 6:e25051.
- [208] Charles W. Morris (1988). *Grundlagen der Zeichentheorie; Ästhetik der Zeichentheorie*. Fischer Verlag.

- [209] Kirsty L. Nash, Craig R. Allen, David G. Angeler, Chris Barichievy, Tarsha Eason, Ahjond S. Garmestani, Nicholas A. J. Graham, Dean Granholm, Melinda Knutson, R. John Nelson, Magnus Nyström, Craig A. Stow, and Shana M. Sundstrom (2014). Discontinuities, cross-scale patterns, and the organization of ecosystems. *Ecology*, 95(3):654–667.
- [210] David M. Needham, Rohan Sachdeva, and Jed A. Fuhrman (2017). Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *The ISME Journal*, 11(7):1614–1629.
- [211] J. A. Nelder and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370.
- [212] Ursula Neumann, Nikita Genze, and Dominik Heider (2017). EFS: an ensemble feature selection tool implemented as r-package and web-application. *BioData Mining*, 10(1).
- [213] Ursula Neumann, Mona Riemenschneider, Jan-Peter Sowa, Theodor Baars, Julia Kälsch, Ali Canbay, and Dominik Heider (2016). Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Mining*, 9(1).
- [214] Felicia N. New and Ilana L. Brito (2020). What is metagenomics teaching us, and what is missed? *Annual Review of Microbiology*, 74(1):117–135.
- [215] Gregoire Nicolis and Ilya Prigogine (1977). *Self-Organization In Nonequilibrium Systems - From Dissipative Structures To Order Through Fluctuations*. John Wiley & Sons Incorporated.
- [216] Søren Nors Nielsen (2016). Second order cybernetics and semiotics in ecological systems—where complexity really begins. *Ecological Modelling*, 319:119–129.
- [217] Viola Nolte, Ram Vinay Pandey, Seffen Jost, Ralph Medinger, Birgit Ottenwälder, Jens Boenigk, and Christian Schlötterer (2010). Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Molecular Ecology*, 19(14):2908–2915.

- [218] A. B. Novikoff (1945). The concept of integrative levels and biology. *Science*, 101(2618):209–215.
- [219] Julia K. Nuy, Matthias Hoetzing, Martin W. Hahn, Daniela Beisser, and Jens Boenigk (2020). Ecological differentiation in two major freshwater bacterial taxa along environmental gradients. *Frontiers in Microbiology*, 11.
- [220] Jay Odenbaugh (2006). The strategy of “the strategy of model building in population biology”. *Biology & Philosophy*, 21(5):607–621.
- [221] Eugene Pleasants Odum (1969). The strategy of ecosystem development. *Science*, 164(3877):262–270.
- [222] Eugene Pleasants Odum (1977). The emergence of ecology as a new integrative discipline. *Science*, 195(4284):1289–1293.
- [223] Eugene Pleasants Odum (1997). *Ecology - A Bridge Between Science And Society*. Sinauer Associates Incorporated.
- [224] Eugene Pleasants Odum (1999). *Ökologie - Grundlagen, Standorte, Anwendung ; 62 Tabellen*. Thieme.
- [225] Jana L. Olefeld, Christina Bock, Manfred Jensen, Janina C. Vogt, Guido Sieber, Dirk Albach, and Jens Boenigk (2020). Centers of endemism of freshwater protists deviate from pattern of taxon richness on a continental scale. *Scientific Reports*, 10(1).
- [226] Angela M. Oliverio, Jean F. Power, Alex Washburne, S. Craig Cary, Matthew B. Stott, and Noah Fierer (2018). The ecology and diversity of microbial eukaryotes in geothermal springs. *The ISME Journal*, 12:1918–1928.
- [227] Maureen A. O’Malley and John Dupré (2005). Fundamental issues in systems biology. *BioEssays*, 27(12):1270–1276.
- [228] Robert V. O’Neill (2001). Is it time to bury the ecosystem concept? (With full military honors, of course!). *Ecology*, 82(12):3275–3284.
- [229] Robert V. O’Neill, Carolyn T. Hunsaker, K. Bruce Jones, Kurt H. Riitters, James D. Wickham, Paul M. Schwartz, Iris A. Goodman, Barbara L. Jackson, and William S. Bail-

- largeon (1997). Monitoring environmental quality at the landscape scale. *BioScience*, 47(8):513–519.
- [230] Anne E. Otwell, Adrián López García de Lomana, Sean M. Gibbons, Mónica V. Orellana, and Nitin S. Baliga (2018). Systems biology approaches towards predictive microbial ecology. *Environmental Microbiology*, 20(12):4197–4209.
- [231] Maureen A. O’Malley (2008). ‘Everything is everywhere: but the environment selects’: ubiquitous distribution and ecological determinism in microbial biogeography. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 39(3):314–325.
- [232] Maureen A. O’Malley and John Dupré (2007). Size doesn’t matter: towards a more inclusive philosophy of biology. *Biology & Philosophy*, 22(2):155–191.
- [233] N. R. Pace (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740.
- [234] Donovan H Parks, Maria Chuvoshina, David W Waite, Christian Rinke, Adam Skarszewski, Pierre-Alain Chaumeil, and Philip Hugenholtz (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10):996–1004.
- [235] Lael Parrott (2010). Measuring ecological complexity. *Ecological Indicators*, 10(6):1069–1076.
- [236] Bernard C. Patten (1978). Systems approach to the concept of environment. *The Ohio Journal of Science*, 78(4):206–222.
- [237] Bernard C. Patten, B.D. Fath, J.S. Choi, S. Bastianoni, S.R. Borrett, S. Brandt-Williams, M. Debeljak, J. Fonseca, W.E. Grant, D. Karnawati, J.C. Marques, A. Moser, F. Müller, C. Pahl-Wostl, R. Seppelt, W.H. Steinborn, and Y.M. Svirezhev (2002). Complex adaptive hierarchical systems. In *Understanding and Solving Environmental Problems in the 21st Century: Toward a New, Integrated Hard Problem Science*, pages 41–94. Elsevier.

- [238] Jan Pawłowski, Laure Apothéloz-Perret-Gentil, and Florian Altermatt (2020). Environmental DNA: What’s behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology*, 29(22):4258–4264.
- [239] John Phillips (1931). The biotic community. *The Journal of Ecology*, 19(1):1.
- [240] Karl Raimund Popper (1990). *A World of Propensities*. Burns & Oates.
- [241] Jean F. Power, Carlo R. Carere, Charles K. Lee, Georgia L. J. Wakerley, David W. Evans, Mathew Button, Duncan White, Melissa D. Climo, Annika M. Hinze, Xochitl C. Morgan, Ian R. McDonald, S. Craig Cary, and Matthew B. Stott (2018). Microbial biogeography of 925 geothermal springs in new zealand. *Nature Communications*, 9(1).
- [242] Ilya Prigogine (1978). Time, structure, and fluctuations. *Science*, 201(4358):777–785.
- [243] Ilya Prigogine and Isabelle Stengers (1981). *Dialog mit der Natur - neue Wege naturwissenschaftlichen Denkens*. Piper.
- [244] Hans Radder (2009). The philosophy of scientific experimentation: a review. *Automated Experimentation*, 1(1):2.
- [245] Kelly S. Ramirez, Christopher G. Knight, Mattias de Hollander, Francis Q. Brearley, Bede Constantinides, Anne Cotton, Si Creer, Thomas W. Crowther, John Davison, Manuel Delgado-Baquerizo, Ellen Dorrepaal, David R. Elliott, Graeme Fox, Robert I. Griffiths, Chris Hale, Kyle Hartman, Ashley Houlden, David L. Jones, Eveline J. Krab, Fernando T. Maestre, Krista L. McGuire, Sylvain Monteux, Caroline H. Orr, Wim H. van der Putten, Ian S. Roberts, David A. Robinson, Jennifer D. Rocca, Jennifer Rowntree, Klaus Schlaeppli, Matthew Shepherd, Brajesh K. Singh, Angela L. Straathof, Jennifer M. Bhatnagar, Cécile Thion, Marcel G. A. van der Heijden, and Franciska T. de Vries (2017). Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nature Microbiology*, 3(2):189–196.
- [246] Maxwell J.D. Ramstead, Axel Constant, Paul B. Badcock, and Karl J. Friston (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*, 31:188–205.
- [247] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

- [248] Marc H V Van Regenmortel (2004). Reductionism and complexity in molecular biology. *EMBO reports*, 5(11):1016–1020.
- [249] Hans-Jorg Rheinberger (1997). Experimental complexity in biology: Some epistemological and historical remarks. *Philosophy of Science*, 64:S245–S254.
- [250] Hans-Jörg Rheinberger (1992). Experiment, difference, and writing: I. tracing protein synthesis. *Studies in History and Philosophy of Science Part A*, 23(2):305–331.
- [251] Hans-Jörg Rheinberger (1992). Experiment, difference, and writing: II. the laboratory production of transfer RNA. *Studies in History and Philosophy of Science Part A*, 23(3):389–422.
- [252] Karine Felix Ribeiro, Leandro Duarte, and Luciane Oliveira Crossetti (2018). Everything is not everywhere: a tale on the biogeography of cyanobacteria. *Hydrobiologia*, 820:23–48.
- [253] Payne Richard J. (2013). Seven reasons why protists make useful bioindicators. *Acta Protozoologica*, 52(3):105–113.
- [254] Jonathan L. Richardson (1980). The organismic community: Resilience of an embattled ecological concept. *BioScience*, 30.
- [255] Thea Van Rossum, Pamela Ferretti, Oleksandr M. Maistrenko, and Peer Bork (2020). Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology*, 18(9):491–506.
- [256] Thea Van Rossum, Michael A. Peabody, Miguel I. Uyaguari-Diaz, Kirby I. Cronin, Michael Chan, Jared R. Slobodan, Matthew J. Nesbitt, Curtis A. Suttle, William W. L. Hsiao, Patrick K. C. Tang, Natalie A. Prystajek, and Fiona S. L. Brinkman (2015). Year-long metagenomic study of river microbiomes across land use and water quality. *Frontiers in Microbiology*, 6:1405.
- [257] Lisa Röttgers and Karoline Faust (2018). Can we predict keystones? *Nature Reviews Microbiology*, 17(3):193–193.
- [258] Melissa A. Rubin and Laura G. Leff (2007). Nutrients and other abiotic factors affecting bacterial communities in an ohio river (USA). *Microbial Ecology*, 54(2):374–383.

- [259] W. G. Rudd, D. C. Herzog, and L. D. Newsom (1984). Hierarchical models of ecosystems. *Environmental Entomology*, 13(2):584–587.
- [260] Michael J. Russell (2017). Life is a verb, not a noun. *Geology*, 45(12):1143–1144.
- [261] Y. Saeys, I. Inza, and P. Larranaga (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.
- [262] Raphael Sagarin and Aníbal Pauchard (2010). Observational approaches in ecology open new ground in a changing world. *Frontiers in Ecology and the Environment*, 8(7):379–386.
- [263] Mark Sagoff (2003). The plaza and the pendulum: two concepts of ecological science. *Biology & Philosophy*, 18(4):529–552.
- [264] Mark Sagoff (2017). On the definition of ecology. *Biological Theory*, 12(2):85–98.
- [265] M. Sagova-Mareckova, J. Boenigk, A. Bouchez, K. Cermakova, T. Chonova, T. Cordier, U. Eisendle, T. Elerseck, S. Fazi, T. Fleituch, L. Frühe, M. Gajdosova, N. Graupner, A. Haegerbaeumer, A.-M. Kelly, J. Kopecky, F. Leese, P. Nöges, S. Orlic, K. Panksep, J. Pawlowski, A. Petrusek, J.J. Piggott, J.C. Rusch, R. Salis, J. Schenk, K. Simek, A. Stovicek, D.A. Strand, M.I. Vasquez, T. Vrålstad, S. Zlatkovic, M. Zupancic, and T. Stoeck (2021). Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. *Water Research*, 191:116767.
- [266] George W. Salt (1979). A comment on the use of the term emergent properties. *The American Naturalist*, 113(1):145–148.
- [267] Stanley N. Salthe (2001). Theoretical biology as an anticipatory text: The relevance of uexküll to current issues in evolutionary systems. *Semiotica*, 2001(134):359–380.
- [268] Stanley N. Salthe (2014). Creating the umwelt: From chance to choice. *Biosemiotics*, 7(3):351–359.
- [269] F. Sanger, S. Nicklen, and A. R. Coulson (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.



- [270] U. Sauer, M. Heinemann, and N. Zamboni (2007). Genetics: Getting closer to the whole picture. *Science*, 316(5824):550–551.
- [271] William M. Schaffer (1981). Ecological abstraction: The consequences of reduced dimensionality in ecological models. *Ecological Monographs*, 51(4):383–401.
- [272] Robert E. Schapire (1990). The strength of weak learnability. *Machine Learning*, 5(2):197–227.
- [273] Patrick D. Schloss (2021). Amplicon sequence variants artificially split bacterial genomes into separate clusters. *Preprint at bioRxiv*.
- [274] Erwin Schrödinger (1944). *What is Life? The Physical Aspect of the Living Cell*. The University Press.
- [275] Jürgen Schwoerbel and Heinz Brendelberger (2005). *Einführung in die Limnologie*. Spektrum Akademischer Verlag.
- [276] Robin Senge, Juan José del Coz, and Eyke Hüllermeier (2013). On the problem of error propagation in classifier chains for multi-label classification. In *Studies in Classification-Data Analysis and Knowledge Organization*, pages 163–170. Springer International Publishing.
- [277] Ashley Shade (2016). Diversity is the question, not the answer. *The ISME Journal*, 11(1):1–6.
- [278] Ashley Shade, Jordan S Read, Nicholas D Youngblut, Noah Fierer, Rob Knight, Timothy K Kratz, Noah R Lottig, Eric E Roden, Emily H Stanley, Jesse Stombaugh, Rachel J Whitaker, Chin H Wu, and Katherine D McMahon (2012). Lake microbial communities are resilient after a whole-ecosystem disturbance. *The ISME Journal*, 6(12):2153–2167.
- [279] Claude E. Shannon (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.
- [280] Claude E. Shannon (1956). The bandwagon. *IRE Transactions on Information Theory*, 2(1):3.

- [281] Daniel Simberloff (1998). Flagships, umbrellas, and keystones: Is single-species management passé in the landscape era? *Biological Conservation*, 83(3):247–257.
- [282] Herbert A. Simon (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482.
- [283] Neil R Smalheiser (2002). Informatics and hypothesis-driven research. *EMBO reports*, 3(8):702–702.
- [284] Lori A. S. Snyder, Nick Loman, Mark J. Pallen, and Charles W. Penn (2008). Next-generation sequencing—the promise and perils of charting the great microbial unknown. *Microbial Ecology*, 57(1):1–3.
- [285] Theodor Sperlea, Stefan Füsler, Jens Boenigk, and Dominik Heider (2018). SEDE-GPS: socio-economic data enrichment based on GPS information. *BMC Bioinformatics*, 19(440).
- [286] Theodor Sperlea, Nico Kreuder, Daniela Beisser, Georges Hattab, Jens Boenigk, and Dominik Heider (2021). Quantification of the covariation of lake microbiomes and environmental variables using a machine learning-based framework. *Molecular Ecology*, 30(9):2131–2144.
- [287] Theodor Sperlea, Lea Muth, Roman Martin, Christoph Weigel, Torsten Waldminghaus, and Dominik Heider (2020). gammaBOriS: Identification and taxonomic classification of origins of replication in Gammaproteobacteria using motif-based machine learning. *Scientific Reports*, 10(1):6727.
- [288] Theodor Sperlea, Jan Philip Schenk, Hagen Dreßler, Daniela Beisser, Georges Hattab, Jens Boenigk, and Dominik Heider (2021). Covariation between the european lake microbiome and surrounding land cover and bioindicator-based insights into the microbiome structure. *In Review at ISME Communications*.
- [289] Will Steffen, Katherine Richardson, Johan Rockström, Hans Joachim Schellnhuber, Opha Pauline Dube, Sébastien Dutreuil, Timothy M. Lenton, and Jane Lubchenco (2020). The emergence and evolution of earth system science. *Nature Reviews Earth & Environment*, 1(1):54–63.

- [290] Lewi Stone and Tom Berman (1993). Positive feedback in aquatic ecosystems: The case of the microbial loop. *Bulletin of Mathematical Biology*, 55(5):919–936.
- [291] G. Sugihara, R. May, H. Ye, C. h. Hsieh, E. Deyle, M. Fogarty, and S. Munch (2012). Detecting causality in complex ecosystems. *Science*, 338(6106):496–500.
- [292] Janusz Szyrmer and Robert E. Ulanowicz (1987). Total flows in ecosystems. *Ecological Modelling*, 35(1-2):123–136.
- [293] Pierre Taberlet, Eric Coissac, François Pompanon, Christian Brochmann, and Eske Willerslev (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8):2045–2050.
- [294] BoonFei Tan, Charmaine Ng, Jean Pierre Nshimiyimana, Lay Leng Loh, Karina Y.-H. Gin, and Janelle R. Thompson (2015). Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges and future opportunities. *Frontiers in Microbiology*, 6.
- [295] A. G. Tansley (1935). The use and abuse of vegetational concepts and terms. *Ecology*, 16(3):284–307.
- [296] Andrew Maltez Thomas and Nicola Segata (2019). Multiple levels of the unknown in microbiome research. *BMC Biology*, 17.
- [297] Luke R. Thompson, Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, Anupriya Tripathi, Sean M. Gibbons, Gail Ackermann, Jose A. Navas-Molina, Stefan Janssen, Evguenia Kopylova, Yoshiki Vázquez-Baeza, Antonio González, James T. Morton, Siavash Mirarab, Zhenjiang Zech Xu, Lingjing Jiang, Mohamed F. Haroon, Jad Kanbar, Qiyun Zhu, Se Jin Song, Tomasz Kosciółek, Nicholas A. Bokulich, Joshua Leffler, Colin J. Brislawn, Gregory Humphrey, Sarah M. Owens, Jarrad Hampton-Marcell, Donna Berg-Lyons, Valerie McKenzie, Noah Fierer, Jed A. Fuhrman, Aaron Clauset, Rick L. Stevens, Ashley Shade, Katherine S. Pollard, Kelly D. Goodwin, Janet K. Jansson, Jack A. Gilbert, Rob Knight, and The Earth Microbiome Project Consortium (2017). A communal catalogue reveals earth’s multiscale microbial diversity. *Nature*, 551:457–463.

- [298] S. G. Tringe (2005). Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557.
- [299] Robert E. Ulanowicz (1986). *Growth and Development*. Springer New York.
- [300] Robert E. Ulanowicz (2019). The tripartite nature of causalities in ecosystem dynamics. *Current Opinion in Systems Biology*, 13:129–135.
- [301] L.A. Urry, M.L. Cain, S.A. Wasserman, P.V. Minorsky, J.B. Reece, and N.A. Campbell (2017). *Campbell Biology*. Pearson Education, Incorporated.
- [302] Antonie van Leeuwenhoek (1677). Observations communicated to the publisher by mr. antony van leewenhoeck in a dutch letter of the 9th octob. 1676. here english'd: concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infused. *Philosophical Transactions of the Royal Society of London*, 12(133):821–831.
- [303] Raf Vanderstraeten (2001). Observing systems: a cybernetic perspective on system/environment relations. *Journal for the Theory of Social Behaviour*, 31(3):297–311.
- [304] Vladimir Vapnik, Steven Golowich, and Alex Smola (1997). Support vector method for function approximation, regression estimation and signal processing. In M. C. Mozer, M. Jordan, and T. Petsche, editors, *Proceedings of the 9th International Conference on Neural Information Processing Systems (NIPS'96)*, pages 281–287. MIT Press.
- [305] Walter Veit (2019). Model pluralism. *Philosophy of the Social Sciences*, 50(2):91–114.
- [306] Vladimir I. Vernadsky (1998). *The Biosphere*. Springer Science & Business Media.
- [307] Joana Amorim Visco, Laure Apothéloz-Perret-Gentil, Arielle Cordonier, Philippe Esling, Loïc Pillet, and Jan Pawlowski (2015). Environmental monitoring: Inferring the diatom index from next-generation sequencing data. *Environmental Science & Technology*, 49(13):7597–7605.
- [308] Heinz von Foerster (2003). *Understanding Understanding - Essays On Cybernetics And Cognition*. Springer Science & Business Media.

- [309] Jakob von Uexküll and Doris L. Mackinnon (transl.) (1926). *Theoretical biology*. K. Paul, Trench, Trubner & co. ltd.; Harcourt, Brace & company, inc London, New York.
- [310] Jakob von Uexküll (1973). *Theoretische Biologie*. Suhrkamp.
- [311] Christine von Weizsäcker and Ernst U. von Weizsäcker (1984). Fehlerfreundlichkeit. In *Offenheit–Zeitlichkeit–Komplexität*, pages 167–201. Campus.
- [312] Ernst U. von Weizsäcker (1974). Erstmaligkeit und Bestätigung als Komponenten der pragmatischen Information. In *Offene Systeme I*. Klett–Cotta, Stuttgart.
- [313] Nicole S. Webster, Michael Wagner, and Andrew P. Negri (2018). Microbial conservation in the Anthropocene. *Environmental Microbiology*, 20(6):1925–1928.
- [314] Stefan Weckx, Roel Van der Meulen, Joke Allemeersch, Geert Huys, Peter Vandamme, Paul Van Hummelen, and Luc De Vuyst (2010). Community dynamics of bacteria in sourdough fermentations as revealed by their metatranscriptome. *Applied and Environmental Microbiology*, 76(16):5402–5408.
- [315] Edward D. Weinberger (2002). A theory of pragmatic information and its application to the quasi-species model of biological evolution. *Biosystems*, 66(3):105–119.
- [316] Edward D. Weinberger (2009). Pragmatic information rates, generalizations of the kelly criterion, and financial market efficiency. *arXiv:0903.2243*, page 0903.2243.
- [317] Marius Welzel, Anja Lange, Dominik Heider, Michael Schwarz, Bernd Freisleben, Manfred Jensen, Jens Boenigk, and Daniela Beisser (2020). Natrix: a Snakemake-based workflow for processing, clustering, and taxonomically assigning amplicon sequencing reads. *BMC Bioinformatics*, 21(526).
- [318] Hans V Westerhoff and Bernhard O Palsson (2004). The evolution of molecular biology into systems biology. *Nature Biotechnology*, 22(10):1249–1252.
- [319] Robert G. Wetzel (2000). *Limnology*. Harcourt Brace College Publishers.
- [320] Norbert Wiener (1961). *Cybernetics - Or, Control and Communication in the Animal and the Machine*. MIT Press.

- [321] Craig E. Williamson, Walter Dodds, Timothy K. Kratz, and Margaret A. Palmer (2008). Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. *Frontiers in Ecology and the Environment*, 6(5):247–254.
- [322] Andrew S. Winston and Daniel J. Blais (1996). What counts as an experiment?: A trans-disciplinary analysis of textbooks, 1930-1970. *The American Journal of Psychology*, 109(4):599.
- [323] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Science.
- [324] C. R. Woese and G. E. Fox (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090.
- [325] C. R. Woese, O. Kandler, and M. L. Wheelis (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579.
- [326] O. Wolkenhauer (2001). Systems biology: The reincarnation of systems theory applied in biology? *Briefings in Bioinformatics*, 2(3):258–270.
- [327] O. Wolkenhauer, M. Mesarović, and P. Wellstead (2007). A plea for more theory in molecular biology. In *Systems Biology. Ernst Schering Research Foundation Workshop*, volume 61. Springer, Berlin, Heidelberg.
- [328] H. Ye and G. Sugihara (2016). Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. *Science*, 353(6302):922–925.
- [329] Hao Ye, Richard J. Beamish, Sarah M. Glaser, Sue C. H. Grant, Chih hao Hsieh, Laura J. Richards, Jon T. Schnute, and George Sugihara (2015). Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences*, 112(13):E1569–E1576.
- [330] Peter Yodzis (1988). The indeterminacy of ecological interactions as perceived through perturbation experiments. *Ecology*, 69(2):508–515.
- [331] Alex E. Yuan and Wenying Shou (2021). Data-driven causal analysis of observational time series in ecology. *Preprint at bioRxiv*.

- [332] Lydia H. Zeglin (2015). Stream microbial diversity in response to environmental changes: review and synthesis of existing research. *Frontiers in Microbiology*, 6.
- [333] Haihan Zhang, Raju Sekar, and Petra M. Visser (2020). Editorial: Microbial ecology in reservoirs and lakes. *Frontiers in Microbiology*, 11.
- [334] Thomas Zoglauer (1996). Can information be naturalized? In K. Kornwachs & K. Jacoby, editor, *Information. New questions to a multidisciplinary concept*, pages 187–207. Berlin: Akademie Verlag.
- [335] Emile Zuckerkandl and Linus Pauling (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8(2):357–366.







## Curriculum vitae

THEODOR SPERLEA

DATE OF BIRTH: 05.04.1992

PLACE OF BIRTH: BIETIGHEIM-BISSINGEN

- since 03/2017 **PhD Student**, Philipps-Universität Marburg.  
Advisor: Prof. Dr. Dominik Heider  
Thesis Title: “The European Lake Microbiome: A Study in Complexity”
- 10/2014 – 02/2017 **Molecular and Cellular Biology (MSc)**, Philipps-Universität Marburg.  
Thesis Advisor: Prof. Dr. Torsten Waldminghaus  
Thesis Title: “Towards an architecture of chromosome maintenance motif patterns on bacterial genomes”
- 10/2011 – 09/2014 **Biology (BSc)**, Philipps-Universität Marburg.  
Thesis Advisor: Prof. Dr. Torsten Waldminghaus  
Thesis Title: “The Role of the DnaA-Box in the Origin of the Secondary Chromosome in *Vibrio cholerae*”

**T**HIS THESIS WAS TYPESET with  $\text{\LaTeX}$ , originally developed by Leslie Lamport and based on Donald Knuth's  $\text{\TeX}$ . The  $\text{\LaTeX}$  source was converted from Pandoc, a Markdown variant developed by John MacFarlane. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. A template that can be used to format a PhD thesis with this look & feel has been released under the permissive MIT (X11) license, and can be retrieved online at [github.com/suchow/](https://github.com/suchow/).